ระบบให้คำแนะนำภาพยนตร์ด้วยวิธีการผสมผสาน

วิทยานิพนธ์

ของ

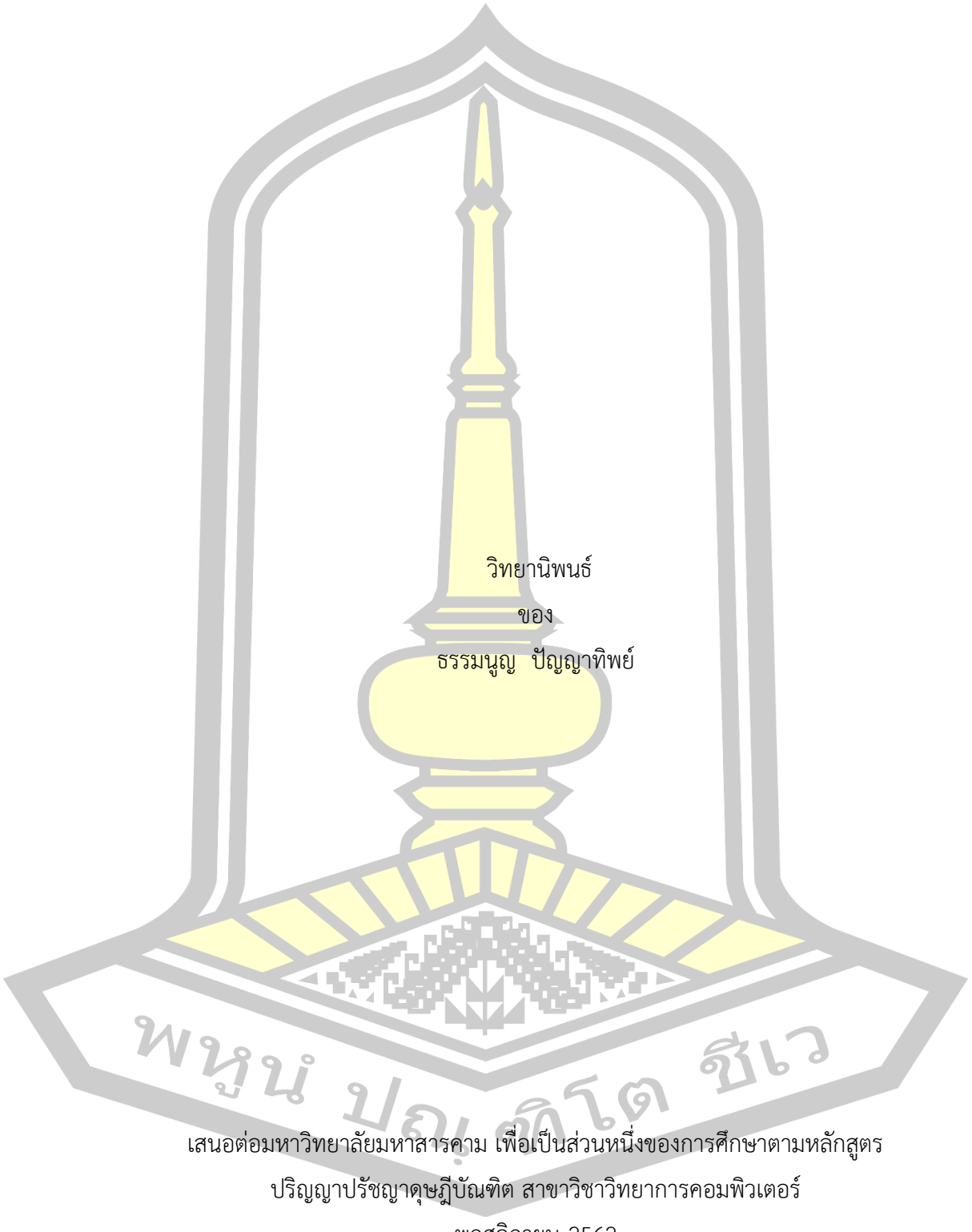
ธรรมนูญ ปัญญาทิพย์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

พฤศจิกายน 2562

ระบบให้คำแนะนำภาพยนตร์ด้วยวิธีการผสมผสาน



วิทยานิพนธ์

ของ

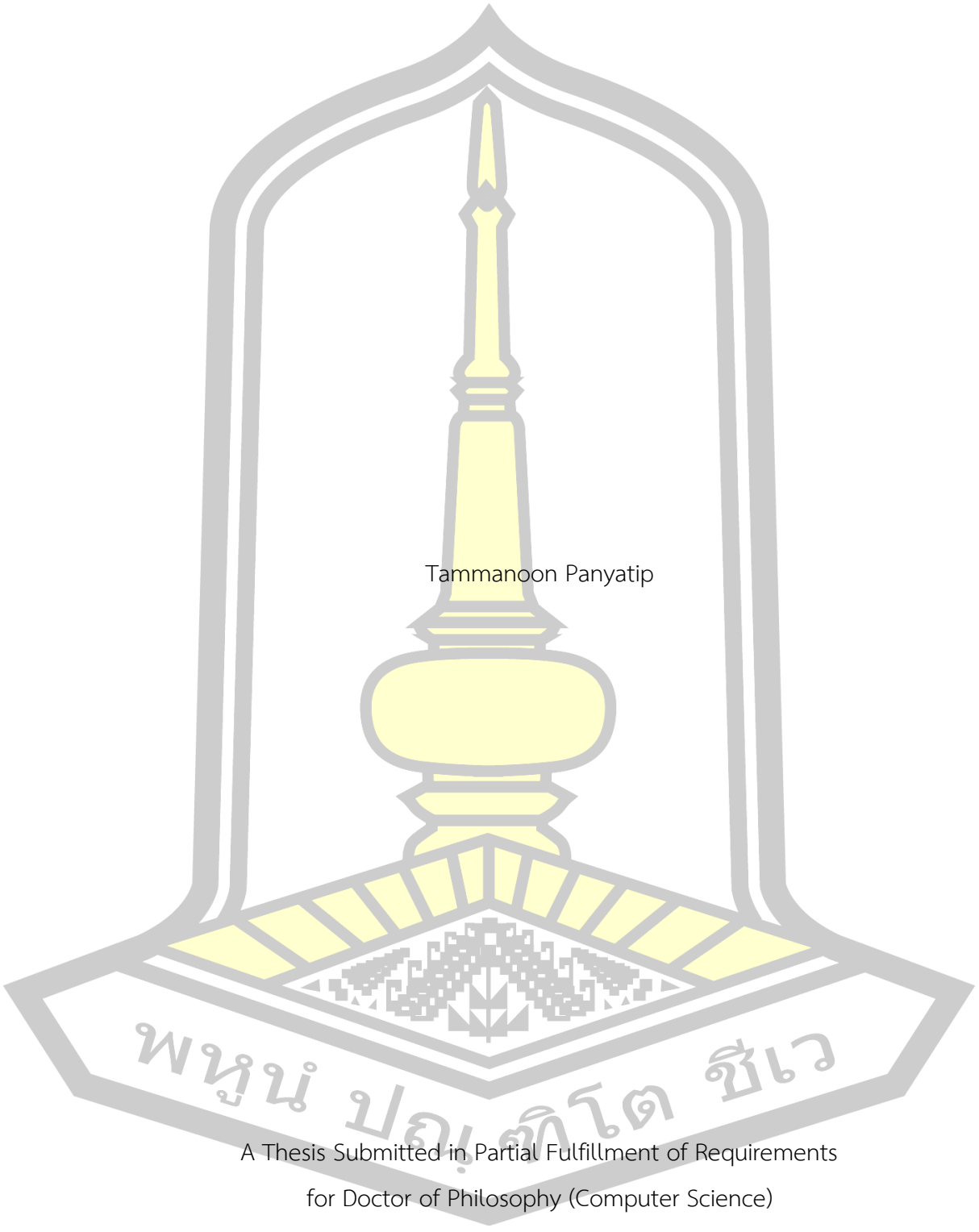
ธรรมนูญ ปัญญาทิพย์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

พฤศจิกายน 2562

Movie Recommendation Using Hybrid Method



Tammanoon Panyatip

A Thesis Submitted in Partial Fulfillment of Requirements

for Doctor of Philosophy (Computer Science)

November 2019

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายธรรมนูญ ปัญญาทิพย์ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผศ. ดร. วรารัตน์ สงฆ์แป้น)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. มนัสวี แก่นอำพรพันธ์)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผศ. ดร. พัฒนพงษ์ ชมภูวิเศษ)

.....กรรมการ
(ผศ. ดร. ฉัตรเกล้า เจริญผล)

.....กรรมการ
(ผศ. ดร. พนิดา ทรงรัมย์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา ปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

.....
(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

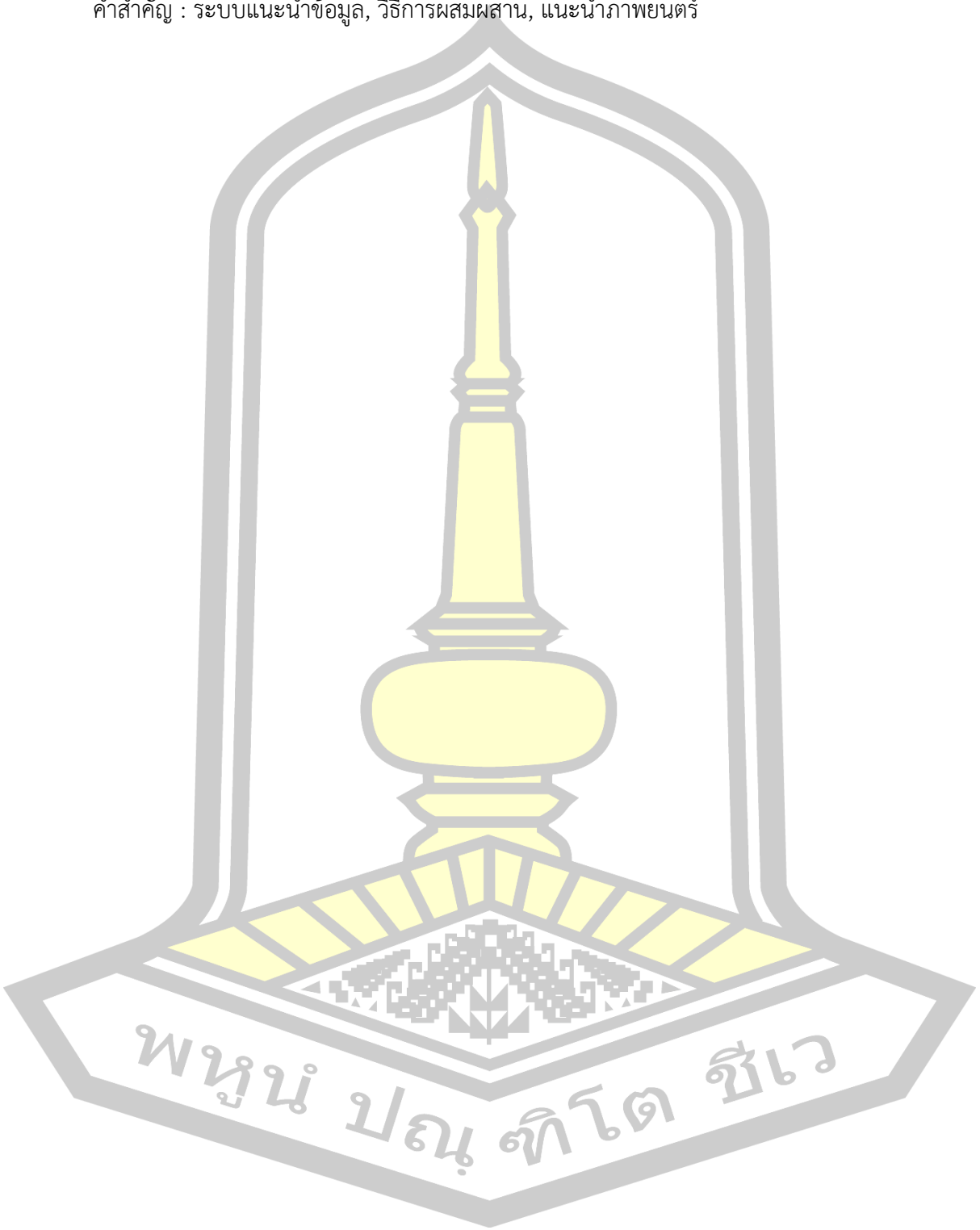
.....
(ผศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง ระบบให้คำแนะนำภาพยนตร์ด้วยวิธีการผสมผสาน
ผู้วิจัย ธรรมนูญ ปัญญาทิพย์
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. มนัสวี แก่นอำพรพันธ์
ผู้ช่วยศาสตราจารย์ ดร. พัฒนพงษ์ ชมภูวิเศษ
ปริญญา ปรัชญาดุษฎีบัณฑิต **สาขาวิชา** วิทยาการคอมพิวเตอร์
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2562

บทคัดย่อ

ระบบแนะนำข้อมูล เป็นเครื่องมือช่วยอำนวยความสะดวกในการตัดสินใจในการค้นหาและค้นคืนข้อมูลเป็นการอาศัยข้อมูลพฤติกรรมความชอบของผู้ใช้ในการช่วยพยากรณ์แนะนำข้อมูล ระบบแนะนำข้อมูลมี 3 ประเภท ได้แก่ วิธีการคัดกรองข้อมูลแบบอิงเนื้อหา วิธีการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมและวิธีการแบบผสมผสาน ซึ่งระบบแนะนำข้อมูลยังมีปัญหาผู้ใช้ใหม่และสินค้าใหม่ งานวิจัยนี้เสนอวิธีการแบบผสมผสานแบบใหม่เพื่อพัฒนากรอบความคิดในการแนะนำภาพยนตร์ด้วยวิธีการผสมผสานที่เกี่ยวข้องกับผู้ใช้ใหม่และข้อมูลภาพยนตร์ใหม่ ข้อมูลที่ใช้ในการทดลองวิจัยประกอบด้วยฐานข้อมูลมูฟวี่เลนส์ (MovieLens) และฐานข้อมูลดิอินเทอร์เน็ตมูวีเดตาเบส (The Internet Movie Database : IMDB) โดยฐานข้อมูล MovieLens เป็นชุดข้อมูลประกอบด้วยข้อมูลจำนวน 100,000 เร็คคอร์ด ผู้ใช้งาน จำนวน 943 คน และภาพยนตร์ จำนวน 1,682 เรื่อง ส่วนฐานข้อมูล IMDB ข้อมูลที่นำมาใช้ประกอบด้วย ข้อมูลนักแสดงชาย นักแสดงหญิง และผู้กำกับ วิธีการดำเนินการวิจัย ด้วยวิธีการแบบผสมผสานแบบรวมคุณลักษณะ จากการคัดกรองข้อมูลแบบอิงเนื้อหา และการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม ลำดับแรก เป็นการเตรียมข้อมูลจากการประมาณค่ากลุ่ม เป็นการคำนวนผลรวมทั้งหมดภายในกลุ่มยกกำลังสองเพื่อให้ได้การจัดกลุ่มที่ดีที่สุด การจัดกลุ่มแบบฟัซซี่ซีมีนเพื่อลดความซับซ้อนของข้อมูลและเพิ่มความเกี่ยวข้องของค่าคะแนนของผู้ใช้กับภาพยนตร์ และการหาค่าความใกล้เคียงด้วยวิธีการ Item based การประมวลผลการแนะนำข้อมูลเป็นการหาค่าเพื่อนบ้านที่ใกล้ที่สุดและรวมน้ำหนัก วัดประสิทธิภาพด้วยค่าความแม่นยำ (Precision) และค่าคลาดเคลื่อนสมบูรณ์เฉลี่ย (Mean Absolute Error : MAE) ผลการวิจัยพบว่า การแก้ปัญหาผู้ใช้ใหม่ ประสิทธิภาพของค่าความแม่นยำร้อยละ 85 ค่าคลาดเคลื่อนสมบูรณ์เฉลี่ย 2.1011 และการแก้ปัญหาข้อมูลใหม่ ประสิทธิภาพของค่าความแม่นยำร้อยละ 87 ค่าคลาดเคลื่อนสมบูรณ์เฉลี่ย 2.0031 โดยสรุป วิธีการแบบผสมผสานที่พัฒนาขึ้น สามารถแนะนำข้อมูลภาพยนตร์ได้อย่างมีประสิทธิภาพ

คำสำคัญ : ระบบแนะนำข้อมูล, วิธีการผสมผสาน, แนะนำภาพยนตร์



TITLE Movie Recommendation Using Hybrid Method

AUTHOR Tammanoon Panyatip

ADVISORS Assistant Professor Manasawee Kaenampornpan , Ph.D.
Assistant Professor Phatthanaphong Chompoowises , Ph.D.

DEGREE Doctor of Philosophy **MAJOR** Computer Science

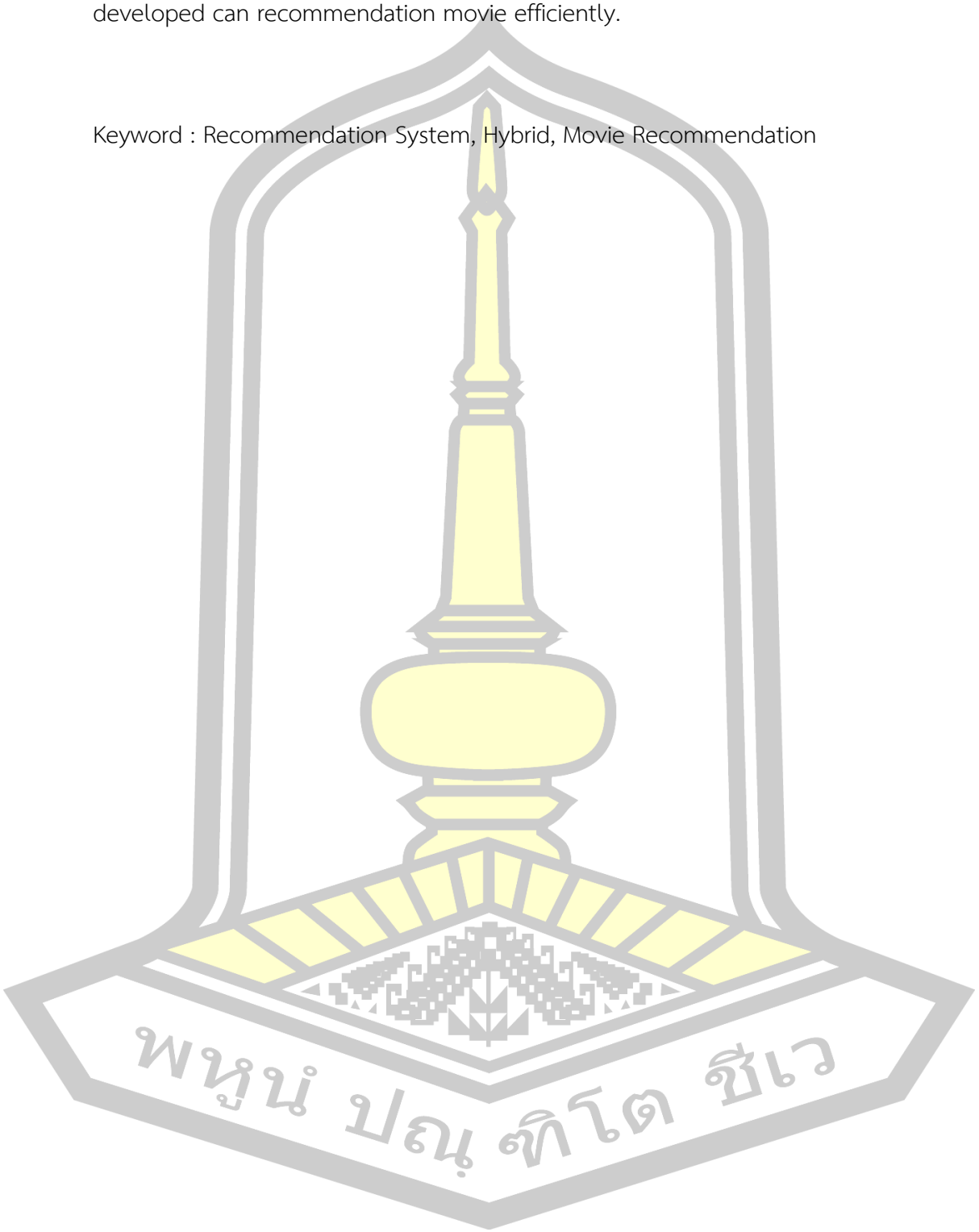
UNIVERSITY Mahasarakham University **YEAR** 2019

ABSTRACT

Recommendation systems are a tool that facilitates decision making during a process of information searching and retrieval. It relies on information of user preference and user behavior in order to recommend the useful information. There are 3 main types of recommendation system, in general. This includes Content Based Filtering (CBF), Collaborative Filtering (CF) and Hybrid Filtering (HF). However, the recommendation system still has problems for new users and new items. This research proposes a new hybrid method to develop the conceptual framework of recommendation system that deals with new user and new movie data. The data used in the research consists of a data from MovieLens and the Internet Movie Database (IMDB). The data from MovieLens contains 100,000 ratings from 943 users on 1,682 movies. As for the IMDB, the data includes information of actors, actresses and directors of the movies. In this work, a hybrid recommendation system with a combination of CBF and CF is introduced. Pre-filtering the data is performed by finding an optimal number of clusters in order to obtain optimal cluster centers. This is done by calculating the total within cluster sum of square. Then the Fuzzy C-mean is implemented in order to reduce the complexity of data and increase the relevance of the user-item ratings. Then the similarity is calculated by using Item Based method. Finally, the K-Nearest Neighbors (KNN) and weight sum of the rating is applied in order to recommend the movies. The performance is measured with precision and Mean Absolute Error (MAE). The research found that for new user data the precision is at 85 percent and MAE value 2.1011. For new item data, the result of research obtains the

precision at 87 percent and MAE value 2.0031. In conclusion, the new hybrid method developed can recommendation movie efficiently.

Keyword : Recommendation System, Hybrid, Movie Recommendation



กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยดี ด้วยความอนุเคราะห์ช่วยชี้แนะอย่างดียิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.มนัสวี แก่นอำพรพันธ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก และผู้ช่วยศาสตราจารย์ ดร.พัฒนพงษ์ ชมภูวิเศษ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่ได้ให้ความกรุณาความเอาใจใส่ดูแล ให้คำปรึกษาและช่วยเหลือมาโดยตลอด จนการทำวิทยานิพนธ์ฉบับนี้สำเร็จตามจุดประสงค์

ขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.วรารัตน์ สงฆ์แป้น ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ฉัตรเกล้า เจริญผล และผู้ช่วยศาสตราจารย์ ดร.พนิดา ทรงรัมย์ กรรมการสอบวิทยานิพนธ์ ที่ให้ข้อเสนอแนะ แนวคิด ตลอดจนแนวทางแก้ไขข้อบกพร่องในการทำวิทยานิพนธ์

ขอกราบขอบพระคุณ คณาจารย์ภาควิชาวิทยาคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ที่อบรมสั่งสอน ช่วยชี้แนะและให้คำปรึกษา ในการศึกษาครั้งนี้

ขอขอบพระคุณผู้บริหาร เพื่อนร่วมงาน คณะอุตสาหกรรมสร้างสรรค์ และคณะวิทยาศาสตร์และเทคโนโลยีสุขภาพ มหาวิทยาลัยกาฬสินธุ์ ที่สนับสนุน คอยช่วยเหลือ และให้กำลังใจแก่ผู้วิจัยมาโดยตลอด

ขอขอบคุณ นางสายทอง ปัญญาทิพย์ นายธนกฤต ปัญญาทิพย์ เด็กชายรฐนนท์ ปัญญาทิพย์ คุณพ่อ คุณแม่ ตลอดจนญาติพี่น้อง รวมทั้งเพื่อนร่วมรุ่น รุ่นพี่ รุ่นน้อง นิสิตปริญญาเอกวิทยาคอมพิวเตอร์ทุกคน ที่ให้ความช่วยเหลือและเป็นกำลังใจด้วยดีมาตลอด

คุณค่าและประโยชน์จากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบเป็นเครื่องบูชาพระคุณบิดา มารดา ครูอาจารย์ และผู้มีพระคุณทุกท่าน ที่ให้ชีวิต ความรัก และความดีงาม จนทำให้ผู้วิจัยประสบผลสำเร็จในการศึกษา

ธรรมนูญ ปัญญาทิพย์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

ความก้าวหน้าในเทคโนโลยีอินเทอร์เน็ตที่ได้มีการเปลี่ยนแปลงเป็นเว็บ 2.0 ซึ่งสามารถติดต่อสื่อสารกันสองทาง และมีผู้ใช้อินเทอร์เน็ตและจำนวนข้อมูลเพิ่มเป็นจำนวนมากใช้กันอย่างแพร่หลาย รวมทั้งช่วยเรื่องการสื่อสารให้มีประสิทธิภาพ สื่อสารได้ในวงกว้าง ได้หลายรูปแบบ เช่น ข้อความ รูปภาพ วิดีโอ เป็นต้น สามารถสื่อสารกับผู้ใช้ที่มีความชื่นชอบในเรื่องเดียวกัน แลกเปลี่ยนความคิดเห็น เข้าถึงกลุ่มเป้าหมายได้รวดเร็วและเป็นช่องทางการสื่อสารได้ตลอดเวลา ดังนั้น จากการเพิ่มขึ้นอย่างรวดเร็วบนของข้อมูลอินเทอร์เน็ต ทำให้การค้นคืนสารสนเทศจากข้อมูลที่มีจำนวนมากให้ตรงกับความชอบของผู้ใช้งานทำได้ยากขึ้นและใช้เวลาเป็นจำนวนมากทำให้เสียเวลาการค้นหา ซึ่งปัจจุบันได้มีระบบแนะนำข้อมูล (Recommender System) สำหรับการกรองข้อมูลมากมาย ในการให้คำแนะนำข้อมูลแก่ผู้ใช้เพื่อให้ตรงกับต้องการ เช่น ภาพยนตร์ เพลง หนังสือ ข่าวสาร และระบบออนไลน์เชิงพาณิชย์อิเล็กทรอนิกส์ ได้นำมาใช้อย่างแพร่หลายและเพิ่มยอดขายสินค้าในระบบจนประสบความสำเร็จ เช่น ร้านขายหนังสือ Amazon.com เว็บไซต์ดูหนัง Netflix.com เป็นต้น

ระบบแนะนำข้อมูล เป็นระบบที่ช่วยกลั่นกรองข้อมูลที่คาดว่าจะตรงกับความชอบของผู้ใช้โดยอาศัยข้อมูลพื้นฐานของประวัติผู้ใช้ กิจกรรมของผู้ใช้ หรือข้อมูลของบุคคลอื่นที่มีความชอบคล้ายคลึงกัน ซึ่งข้อมูลที่นำมาใช้นั้นมีลักษณะคือแบบชัดเจน (Explicit) เช่น ค่าคะแนน และแบบไม่ชัดเจน (Implicit) เช่น ข้อมูลพฤติกรรมของผู้ใช้ [2] [3] ทั้งนี้ ระบบแนะนำข้อมูล [1] [4] ได้แก่ 1) วิธีการคัดกรองข้อมูลแบบอิงเนื้อหา (Content-Based Filtering : CBF) เป็นวิธีการนำพฤติกรรมของผู้ใช้หรือลูกค้ามาคัดกรองข้อมูลคุณสมบัติของรายการหรือสินค้าและประวัติของผู้ใช้หรือลูกค้า แต่ยังมีปัญหาเกี่ยวกับผู้ใช้ใหม่ (New User) ที่มีข้อมูลการจากพฤติกรรมความชอบต่อชิ้นสินค้ายังไม่มากพอที่จะนำไปใช้ในการแนะนำข้อมูลและการดึงคุณสมบัติของสินค้าออกมาใช้ให้เหมาะสม และในการแนะนำด้วยเทคนิคการคัดกรองข้อมูลแบบอิงเนื้อหานั้น ระบบจะไม่สามารถแนะนำสิ่งที่แตกต่างไปจากสิ่งที่ผู้ใช้เคยใช้งานหรือมีประสบการณ์กับชิ้นข้อมูลนั้นมาก่อน ทำให้ไม่สามารถแนะนำสิ่งที่แตกต่างไปจากสิ่งที่ผู้ใช้งานมีความชอบได้ 2) วิธีการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering : CF) เป็นการคัดกรองข้อมูลโดยอาศัยข้อมูลพฤติกรรมในอดีตในช่วยใน

การแนะนำข้อมูล เช่น การซื้อสินค้า คะแนนค่าคะแนน การค้นหาค่าความใกล้เคียงกันของผู้ใช้หรือรายการสินค้าและใช้ข้อมูลนี้เพื่อที่จะหารายการที่น่าสนใจแนะนำให้กับกับผู้ใช้ [5] ซึ่งวิธีการนี้ยังมีปัญหาข้อมูลสินค้าใหม่ (New Item) ที่ยังไม่มีคะแนนค่าคะแนน และคะแนนค่าคะแนนเบาบาง (Sparsity) ที่นำมาใช้ในการคำนวณค่าความใกล้เคียง ผู้ใช้งานที่อยู่ในระบบมีจำนวนไม่เพียงพอหรือน้อยมาก จึงไม่สามารถทำการจับคู่ ความคล้ายคลึงกับผู้ใช้งานอื่น ๆ ได้ ทำให้ผลลัพธ์ในการแนะนำข้อมูลมีคุณภาพต่ำ 3) การคัดกรองข้อมูลผู้ใช้ เป็นการนำข้อมูลของผู้ใช้มาทำการคัดกรองในการแก้ไขปัญหาผู้ใช้ใหม่ และ 4) วิธีการคัดกรองข้อมูลแบบผสมผสาน (Hybrid Filtering : HF) สามารถแก้ไขปัญหา ดังกล่าวข้างต้นได้

นอกจากนี้ ปัจจุบันข้อมูลได้มีขนาดเพิ่มมากขึ้น ใช้ตัวแปรที่มีจำนวนมาก ทำให้ข้อมูลกระจายตัวมากเกินไปและมีข้อมูลหลายมิติ (Multi Dimension) ส่งผลให้ใช้เวลาในการประมวลผลมากและประสิทธิภาพในการแนะนำข้อมูลลดลง ซึ่งได้มีนักวิจัยได้นำเสนอการแก้ไขปัญหาดังกล่าวหลายวิธีที่แตกต่างกัน เช่น การจัดเตรียมข้อมูล (Data Preparation) ด้วยการแยกตัวแปรคำนวณ [6] การเตรียมข้อมูลด้วยการจำแนก (Classification) ข้อมูลบริบทและไม่ใช่บริบทก่อนเข้าสู่กระบวนการประมวลผล [7] การแบ่งกลุ่มข้อมูล (Clustering) [8] เป็นต้น ซึ่งในกระบวนการเตรียมข้อมูลด้วยวิธีการแบ่งกลุ่มข้อมูลนั้นเป็นการจัดให้ข้อมูลที่มีความคลายคลึงกันอยู่ในกลุ่มเดียวกัน ด้วยวิธีการดังกล่าวยังมีปัญหา ไม่ว่าจะเป็นการกำหนดกลุ่มข้อมูลด้วยการระบุค่า k ไม่เหมาะสม ทำให้กระบวนการจัดกลุ่มข้อมูลใช้เวลานานและไม่มีคุณภาพ งานวิจัยนี้ ผู้วิจัยจึงมีแนวคิดพัฒนาขั้นตอนวิธีการแก้ปัญหาดังกล่าว ก่อนที่จะนำข้อมูลเข้าสู่กระบวนการประมวลผลการแนะนำข้อมูล ซึ่งจะช่วยเพิ่มประสิทธิภาพการคัดกรองแบบผสมผสานให้มีประสิทธิภาพมากยิ่งขึ้น

1.2 วัตถุประสงค์ของการวิจัย

เพื่อพัฒนากรอบความคิดในการแนะนำภาพยนตร์ด้วยวิธีการผสมผสาน

1.3 ความสำคัญของการวิจัย

จากการที่มีผู้ใช้เว็บไซต์มากขึ้นทำให้เกิดปัญหาการขยายตัวของข้อมูล ข้อมูลมีขนาดเพิ่มมากขึ้น การกระจายตัวมากเกินไปและมีข้อมูลหลายมิติ ส่งผลให้ระบบแนะนำข้อมูลมีความล่าช้าไม่ตรงกับความต้องการ วิธีการเตรียมข้อมูล การจำแนกข้อมูลด้วยวิธีการจัดกลุ่ม การหาค่าความใกล้เคียง ก่อนเข้าสู่กระบวนการประมวลผลแนะนำข้อมูลเป็นอีกวิธีการหนึ่งในการเพิ่มประสิทธิภาพ

ของระบบแนะนำข้อมูลซึ่งการกำหนดขนาดของกลุ่มให้เหมาะสมกับข้อมูล การคัดกรองข้อมูลที่เหมาะสมกับข้อมูล ซึ่งจะช่วยเพิ่มประสิทธิภาพการคัดกรองแบบผสมผสานให้มีประสิทธิภาพมากยิ่งขึ้น

1.4 ขอบเขตของการวิจัย

1. ข้อมูลที่ใช้ในงานวิจัยนี้ เป็นฐานข้อมูลการแนะนำข้อมูลภาพยนตร์ จากฐานข้อมูล MovieLens และ IMDB ที่มีเรื่องของภาพยนตร์ตรงกับข้อมูล MovieLens ชุดข้อมูลประกอบด้วยข้อมูลค่าคะแนน 100,000 เร็คคอร์ด จากการเก็บรวบรวมข้อมูลผู้ใช้งาน จำนวน 943 คน และภาพยนตร์จำนวน 1,682 เรื่อง

2. พัฒนาขั้นตอนการจัดกลุ่มข้อมูล และจำแนกข้อมูล ก่อนเข้าสู่กระบวนการประมวลผลการแนะนำข้อมูล

3. พัฒนาระบบแนะนำข้อมูลด้วยวิธีการแบบผสมผสาน

1.5 นิยามศัพท์เฉพาะ

1. ค่าคะแนน (Rating) หมายถึง ข้อมูลที่ผู้ใช้ให้คะแนนชอบความในการดูภาพยนต์ โดยจะมีคะแนนระหว่าง 1-5

2. ข้อมูลเบาบาง (Sparsity) หมายถึง ค่าคะแนนที่ผู้ใช้ให้คะแนนความชอบในภาพยนตร์แต่ละเรื่องมีจำนวนน้อย

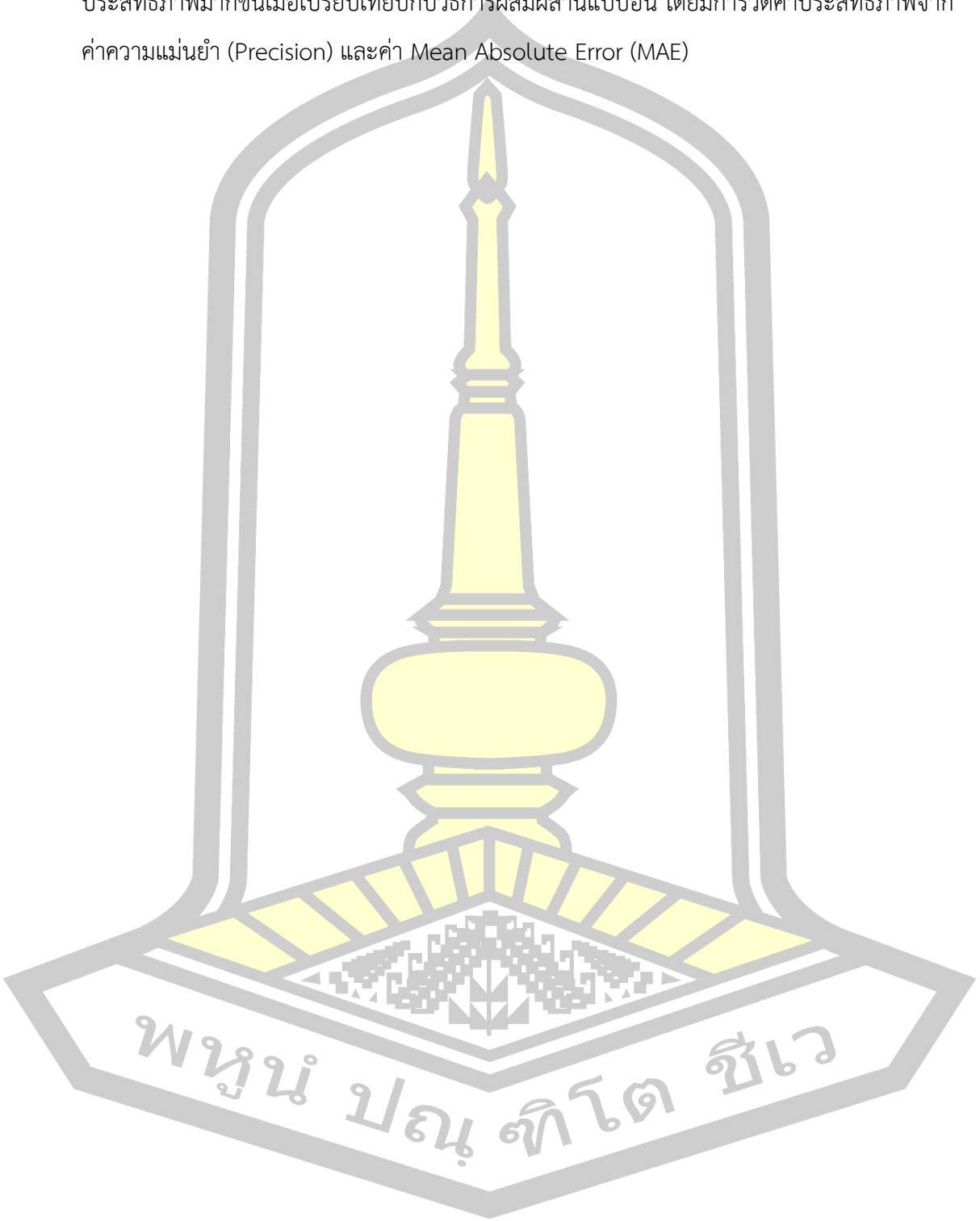
3. ผู้ใช้ใหม่ (New User) หมายถึง ผู้ใช้หรือลูกค้า ที่พึ่งสมัครสมาชิก ยังไม่มีประวัติในการดูภาพยนตร์

4. สินค้าใหม่ (New Item) หมายถึง ภาพยนตร์ที่มาใหม่ ทำให้ยังไม่มีค่าคะแนน

5. การแนะนำข้อมูล (Recommender) หมายถึง การแนะนำข้อมูลภาพยนตร์ โดยอาศัยเทคนิควิธีการคัดกรองแบบผสมผสาน

6. วิธีการผสมผสาน (Hybrid) หมายถึง การจัดกลุ่มข้อมูลที่มีคุณลักษณะใกล้เคียงกันของการคัดกรองข้อมูลแบบอิงเนื้อหา ประกอบด้วยข้อมูลเพศ อายุ อาชีพ ที่มีความสัมพันธ์กับลักษณะกับการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมประกอบด้วย รหัสผู้ใช้ รหัสภาพยนตร์ ค่าคะแนน เพื่อช่วยเสริมหรือปรับปรุงอีกวิธีการหนึ่งให้ระบบแนะนำข้อมูลมีความถูกต้องมากขึ้น

7. การเพิ่มประสิทธิภาพ หมายถึง วิธีการผสมผสานที่ทำให้ระบบแนะนำข้อมูลมีประสิทธิภาพมากขึ้นเมื่อเปรียบเทียบกับวิธีการผสมผสานแบบอื่น โดยมีการวัดค่าประสิทธิภาพจากค่าความแม่นยำ (Precision) และค่า Mean Absolute Error (MAE)

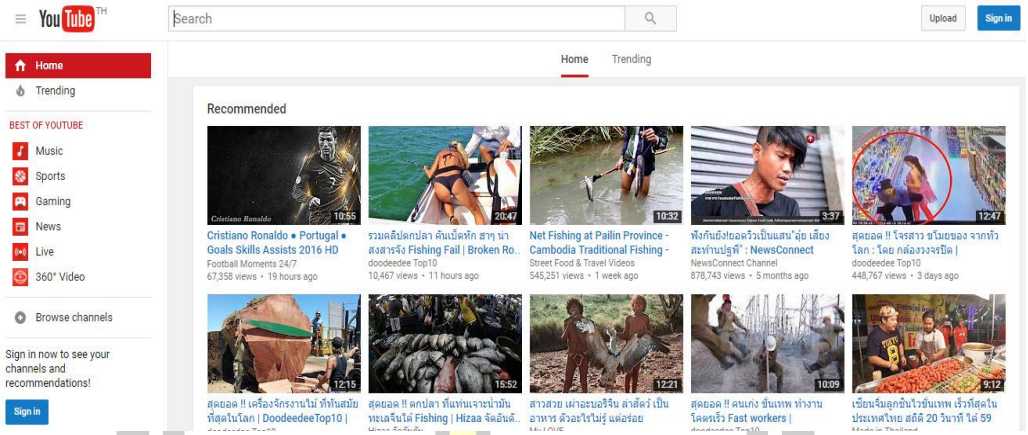


บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การวิจัยเรื่อง การเพิ่มประสิทธิภาพระบบให้คำแนะนำข้อมูลด้วยวิธีการผสานผสาน มีทฤษฎีและงานวิจัยที่เกี่ยวข้อง ได้แก่ ระบบแนะนำข้อมูล แหล่งข้อมูล การจัดเตรียมข้อมูลสำหรับระบบแนะนำข้อมูล ประเภทของระบบการประมวลข้อมูล การวัดประสิทธิภาพ และงานวิจัยที่เกี่ยวข้อง รายละเอียดดังนี้

2.1 ระบบแนะนำข้อมูล

ระบบแนะนำข้อมูล เป็นการนำเสนอประสบการณ์เฉพาะบุคคลให้เกิดขึ้นกับลูกค้า ด้วยการเสนอข้อมูลข่าวสารที่ถูกกลั่นกรองเฉพาะสำหรับลูกค้าแต่ละคน ผ่านช่องทางการติดต่อสื่อสารหลาย ๆ ช่องทาง จะเป็นการเพิ่มโอกาสที่ข้อมูลจะไปถึงลูกค้าได้มากยิ่งขึ้น และทำให้ลูกค้าสามารถนำข้อมูลเหล่านั้นไปใช้ประโยชน์ในทางปฏิบัติได้อย่างรวดเร็ว [9] ระบบแนะนำเป็นเครื่องมือหนึ่งของการทำการตลาดลูกค้าสัมพันธ์อิเล็กทรอนิกส์ (Electronic Customer Relationship Marketing: eCRM) และถูกใช้ในการประมาณการหรือสร้างตัวแบบในการเลือกบริโภคสินค้าในแง่ทางการตลาด ต่อมาเริ่มมีการขยายการวิจัยให้อิสระมากขึ้น กลางปี ค.ศ. 1990 นักวิจัยมุ่งวิจัยไปที่ปัญหาของโครงสร้างของค่าคะแนน ทำให้ปัญหาของการวิจัยถูกลดเป็นปัญหาการประมาณค่า ค่าคะแนน ให้กับชิ้นข้อมูลที่ผู้ใช้ยังไม่เคยประสบการณ์มาก่อน นอกจากนี้แล้วระบบแนะนำยังมีประเด็นอื่น ในการวิจัยหรือแก้ปัญหาได้แก่ บางระบบมีชิ้นข้อมูลจำนวนมากหรือมีผู้ใช้งานจำนวนมาก ในระบบแนะนำข้อมูลโอกาสที่ชิ้นข้อมูลแต่ละชิ้นจะถูกแนะนำขึ้นอยู่กับตัวชี้วัดของมัน ซึ่งก็คือค่าคะแนนนั่นเอง โดยจะบอกว่าผู้ใช้ชอบชิ้นข้อมูลนี้มากน้อยเท่าไหร่อย่างไรก็ดี Xuan Nhat Lam และคนอื่น [10] และ Datta และคนอื่น [11] ได้กล่าวว่า สามารถเพิ่มตัวชี้วัดอื่น ๆ เข้าไปได้ด้วยได้แก่ คุณลักษณะของผู้ใช้แต่ละคน เช่น อายุ เพศ การศึกษา รายได้ เป็นต้น และ Ghazanfar และ Bennett [12] ได้ คุณลักษณะของชิ้นข้อมูลแต่ละชิ้น เช่น ถ้าเป็นข้อมูลภาพยนตร์ คุณลักษณะได้แก่ ชื่อเรื่อง ประเภทหนัง ผู้กำกับ นักแสดงนำ เป็นต้น นั่นหมายความว่า ความสามารถในการแนะนำถูกแทนด้วยค่าคะแนนกับข้อมูลเริ่มต้นของระบบที่เกี่ยวกับผู้ใช้หรือค่าคะแนนที่เคยให้ไว้ [13] เช่น การแนะนำวิดีโอจากเว็บไซต์ยูทูป ดังรูปที่ 1



รูปที่ 1 ตัวอย่างการแนะนำวิดิโอในเว็บไซต์

ส่วนประกอบของระบบแนะนำข้อมูล โดยทั่วไประบบแนะนำข้อมูล มีดังนี้

2.1.1 ส่วนของอินพุตข้อมูล ประกอบด้วย

1) ข้อมูลนำเข้า ที่ประกอบไปด้วยข้อมูลผู้ใช้ คุณลักษณะของข้อมูล ข้อมูลคะแนน ค่าคะแนน และข้อมูลบริบทอื่นๆ ที่เกี่ยวข้อง เป็นต้น

2) การเตรียมข้อมูล เป็นการจัดการข้อมูลก่อนการประมวลผล เพื่อให้ได้ข้อมูลที่มีประสิทธิภาพ เช่น การหาความถี่ของข้อมูล การสกัดคุณลักษณะของข้อมูล การลดมิติของข้อมูล การจัดกลุ่ม การจำแนกประเภท เป็นต้น

2.1.2 ส่วนของการประมวลผลระบบแนะนำข้อมูล ประกอบด้วย 3 วิธี ดังนี้

1) การคัดกรองข้อมูลแบบอิงเนื้อหา (Content-Based Filtering : CBF) ผู้ใช้ระบบจะได้รับการแนะนำชิ้นข้อมูลหรือสินค้าที่มีความคล้ายคลึงกับชิ้นข้อมูลหรือสินค้าที่ผู้ใช้เคยมีประสบการณ์ที่ดีมาแล้วในอดีต เทคนิคที่เคยใช้ในงานวิจัย เช่น การจัดกลุ่ม (Clustering) การจำแนกประเภทข้อความ (Text Classification) การหาความใกล้เคียง (Similarity) เป็นต้น

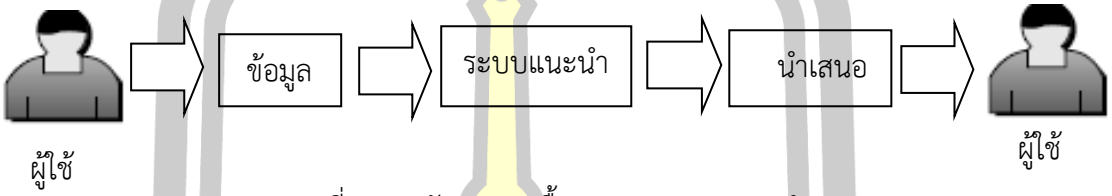
2) การคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering : CF) ผู้ใช้จะได้รับการแนะนำชิ้นข้อมูลหรือสินค้าที่ผู้ใช้ท่านอื่นที่มีรสนิยมคล้ายกันเคยมีประสบการณ์ที่ดีกับชิ้นข้อมูลนั้นในอดีต เทคนิคที่เคยใช้ในงานวิจัยได้แก่ การหาค่าความใกล้เคียง (Similarity) การจัดกลุ่ม (Clustering) เป็นต้น

3) การคัดกรองข้อมูลแบบผสม (Hybrid Filtering : HF) เป็นการผสมผสานระหว่างการคัดกรองข้อมูลแบบอิงเนื้อหากับการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม

2.1.3 ส่วนของการแสดงผล เป็นส่วนการแนะนำข้อมูล มีรูปแบบอยู่ 2 รูปแบบคือ

1) Top N Recommendation โดยจะนำเสนอชิ้นข้อมูล N ชิ้นที่ตรงกับความต้องการของผู้ใช้มากที่สุด

2) Predicted Value ระบบจะทำเสนอชิ้นข้อมูลพร้อมทั้งแสดงข้อมูลค่าคะแนนที่ระบบได้ทำนายเอาไว้



รูปที่ 2 สถาปัตยกรรมพื้นฐานของระบบแนะนำ

จากรูปที่ 2 สถาปัตยกรรมพื้นฐานของระบบแนะนำ เป็นระบบให้การแนะนำชิ้นข้อมูลให้กับผู้ใช้โดยที่ระบบจะทำการนำเสนอชิ้นข้อมูลที่คาดว่าผู้ใช้จะให้ความสนใจ หรือ คาดว่าเป็นชิ้นข้อมูลที่ผู้ใช้ต้องการหรือมองหาอยู่ ซึ่งเป็นการสร้างความพึงพอใจแก่ผู้ใช้ทางอ้อมอีกวิธีหนึ่ง สำหรับการทำนายค่าความพึงพอใจนั้นสามารถคำนวณได้จากคะแนนที่ผู้ใช้ได้เคยให้คะแนนกับชิ้นข้อมูลต่าง ๆ ซึ่งเรียกว่า ค่าคะแนน ซึ่งสามารถนำไปคำนวณร่วมกับค่าคะแนน ของผู้ใช้ที่มีความมีความคล้ายคลึงกันได้ จากนั้นนำค่าความพึงพอใจของผู้ใช้ที่มีความคล้ายคลึงกันมาเป็นส่วนร่วมในการพิจารณาค่าความพึงพอใจของผู้ใช้แต่ละคนแตกต่างไปตามรสนิยมของผู้ใช้แต่ละคนนั้น เมื่อทำนายค่าเสร็จแล้วระบบจะทำการนำเสนอผลการแนะนำออกมา

ดังนั้นในการทำระบบแนะนำข้อมูล จึงมีขั้นตอนการทำงานในภาพรวม คือ แหล่งข้อมูล การเตรียมข้อมูล ประเภทของระบบแนะนำข้อมูล และการวัดประสิทธิภาพ รายละเอียดดังนี้

2.2 การกลั่นกรองข้อมูลสำหรับระบบแนะนำข้อมูล

การกลั่นกรองข้อมูลก่อนเข้าสู่การประมวลของแนะนำข้อมูลนั้นถือว่าเป็นสิ่งสำคัญ ไม่ว่าจะเป็นการทำความสะอาดข้อมูล (Data Cleansing) ซึ่งจะเป็นการลบหรือลดข้อมูลที่เป็นสิ่งรบกวนหรือการจัดการกับข้อมูลที่ขาดหายไปด้วยวิธีการแทนค่าข้อมูลที่มีค่าเชิงสถิติ ความเกี่ยวข้องเนื่องของข้อมูล (Relevance Analysis) เป็นการตรวจสอบแอตทริบิวต์ที่เกี่ยวเนื่องหรือซ้ำซ้อนกันในการหาค่า

ความใกล้เคียงหรือแตกต่างกันด้วยการพิจารณาแอตทริบิวต์ทีละคู่ แล้วทำการเลือกแอตทริบิวต์ที่น่าสนใจในการลดปริมาณแอตทริบิวต์และเป็นการเพิ่มค่าความถูกต้องของผลลัพธ์ได้ การเปลี่ยนแปลงหรือเปลี่ยนรูปข้อมูลและลดจำนวนข้อมูล (Data Transformation and Reduction) ช่วงข้อมูลที่อินพุตต์เข้ามาหรือค่าข้อมูลที่มีระยะห่างมาก อาจทำการเปลี่ยนแปลงหรือเปลี่ยนรูปด้วยวิธีการ Normalization ที่จะทำการปรับเปลี่ยนค่าในแอตทริบิวต์ให้อยู่ในช่วงที่กำหนด เช่น ช่วง 0 ถึง 1 หรือช่วง -1 ถึง 1 เป็นต้น ซึ่งจะช่วยในการหาความแตกต่างระหว่างข้อมูล และสามารถช่วยลดจำนวนข้อมูลที่ต้องการทำการพิจารณาได้ด้วยการประยุกต์ใช้เทคนิคต่าง ๆ ด้วยวิธีการดังนี้

2.2.1 การจำแนกประเภทข้อมูลด้วยวิธี K-Nearest Neighbors (KNN)

หลักการทำงานของ KNN เป็นการวัดระยะห่างระหว่างข้อมูลที่ต้องการพยากรณ์กับข้อมูลที่อยู่ใกล้เคียง โดยกำหนดจำนวนเป็น K ตัว ปกติวิธีการวัดระยะห่างจะใช้แบบยูคลิเดียน (Euclidean) ดังสมการ (1)

$$D_{Euclidean} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_L - y_L)^2} \quad (1)$$

โดยที่ x_1 คือ แอตทริบิวต์ที่ 1 ของข้อมูลชุดที่ 1

y_1 คือ แอตทริบิวต์ที่ 1 ของข้อมูลชุดที่ 2

โดยข้อมูลทั้งสองตัว (x และ y) มีจำนวนแอตทริบิวต์เท่ากับ L

สำหรับงานวิจัยที่นำวิธีการ KNN มาประยุกต์ใช้กับระบบแนะนำข้อมูล ซึ่งได้นำวิธีการของ KNN อย่างเช่น Baltrunas และ Ricci [20] ได้ทำการแยกตารางการคำนวณจากข้อมูลค่าคะแนน ด้วยวิธีการคัดกรองแบบพึ่งพาผู้ใช้ร่วม ทำการจำแนกข้อมูลด้วยวิธีการเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest neighbors (KNN) การแยกตารางคำนวณ หาค่าเฉลี่ย และทำการเปรียบเทียบกับวิธีการการแยกตารางข้อมูลสินค้าในการคำนวณ (Item-split) Campos และคณะ [21] ได้ทำการศึกษาระบบแนะนำข้อมูลภาพยนตร์ด้วยบริบทในการเปรียบเทียบวิธีการ Pre-filtering Post-filtering และโมเดลบริบท ในการแก้ไขปัญหามัลติเลเบลและข้อมูลหลายมิติ ฐานข้อมูลเป็นภาพยนตร์ มีข้อมูลประเภทภาพยนตร์ ข้อมูลค่าคะแนน ข้อมูลเวลาประกอบด้วยวันปกติประกอบด้วย เช้า บ่าย กลางคืน ไม่ระบุ วันในสัปดาห์ ประกอบด้วย วันทำงาน วันหยุด ไม่ระบุ ข้อมูลเพื่อน ประกอบด้วย คนเดียว กับสามีหรือภรรยา ครอบครัว เพื่อน ไม่ระบุ ข้อมูลพวกนี้จะเพิ่มด้วยค่าค่าคะแนนที่ผู้ใช้มีความชอบต่อภาพยนตร์ คือ ค่าคะแนน 1-5 วิธีการ Pre-filtering คำนวณด้วยการแยกรายการ ด้วย KNN วิธีการ Post-filtering และโมเดลบริบท โดยใช้การจำแนกประเภทด้วย นาอีฟเบย์ Random Forest MLP SVM ผลปรากฏว่าการแยกตารางคำนวณที่ใช้กับ Pre-

filtering มีประสิทธิภาพมากที่สุด และ Zheng Mobasher และคณะ [22] ได้ศึกษาบริบทระบบแนะนำใช้มัลติเลเบลแยกประเภทข้อมูล ทดลองใช้กับข้อมูล AdomMovie LDOS-CoMoDa และ TripAdviser ทำการ Pre-filtering ด้วยเทคนิค Bag of words ด้วยการเลือกคุณลักษณะของข้อมูลโดยการพิจารณาบริบทที่เกี่ยวข้อง ทำการทดลองด้วยการแยกประเภท เปรียบเทียบผลระหว่าง KNN ต้นไม้ตัดสินใจ J48 นาอีฟเบย์ เบย์เซียนเน็ตเวิร์ค และ SVM ผลปรากฏว่าวิธีการ SVM มีประสิทธิภาพมากที่สุด

2.2.2 การประมาณค่ากลุ่ม (k) ซึ่งก่อนจะทำการจัดกลุ่ม ต้องการค่า k ของกลุ่มก่อน ซึ่งผู้วิจัยได้ทำการค้นหาค่า k ด้วยการหาค่าระยะห่างของข้อมูลด้วยกำหนดเป็นกราฟเส้น เพื่อดูความหักเหของเส้นกราฟ เมื่อได้ค่า k จึงทำการจัดกลุ่มของผู้ใช้ หรือข้อมูลลักษณะของภาพยนตร์ สามารถทำได้ดังนี้

1) คำนวนการจัดกลุ่ม เช่น ฟัซซี่ซีมีน หรือ เคมีน แล้วเพื่อให้เห็นความแตกต่างระหว่างค่า k ที่กำหนด เช่น กำหนดค่า k=10

2) สำหรับค่า k เป็นการคำนวนผลรวมทั้งหมดภายในกลุ่มยกกำลังสอง คือการวัดความแปรปรวนของข้อมูลภายในกลุ่ม

3) สร้างกราฟเส้น จากการคำนวนในข้อ 2

4) กราฟเส้นจะมีจุดหักเห ซึ่งเป็นจุดที่ชี้ให้เห็นความแตกต่างในการจัดกลุ่ม

2.2.3 ฟัซซี่ซีมีน (Fuzzy C-means : FCM) เป็นอัลกอริทึมที่ยอมให้ข้อมูลในแต่ละคลัสเตอร์มีการ ซ้อนทับกันหรือซ้ำกันได้ วิธีการนี้เป็นการจัดกลุ่มที่มีใช้อย่างแพร่หลายในงานด้านต่างๆ เช่น การแพทย์ วิทยาศาสตร์ วิศวกรรมศาสตร์ โดยอาศัยการให้ค่าการเป็นสมาชิกของข้อมูลต่อ กลุ่มข้อมูลต่างๆ การได้มาซึ่งค่าการเป็นสมาชิกส่วนหนึ่งมาจากการวัดระยะทางระหว่างข้อมูล และจุดศูนย์กลางของกลุ่มเหล่านั้น การวัดระยะทางจึงมีความสำคัญต่อการจัดกลุ่ม ซึ่งวิธีการ วัดระยะทางนั้นมีหลายวิธีการอาจเป็นการวัดระยะทางแบบยูคลิเดียน (Eulidence Distance) หรือการวัดระยะทางแบบมหาลาโนบิส (Mahalanobis Distance) สำหรับการวัดระยะทางแบบ ยูคลีเดียนนั้นไม่เหมาะกับข้อมูลที่เกี่ยวเนื่องกัน สำหรับการวัดระยะทางแบบมหาลาโนบิสนั้น เหมาะสำหรับกลุ่มข้อมูลที่มีข้อมูลโดดออกจากกลุ่ม (Outlier) และกลุ่มข้อมูลที่มีข้อมูลหนาแน่นต่างๆ

การจัดกลุ่มแบบฟัซซี่นั้นเป็นเทคนิคในการจัดกลุ่มที่แก้ไขข้อเสียของเคมีน เนื่องจากเคมีนไม่เหมาะกับข้อมูลที่มีความสัมพันธ์กัน (Correlation) เนื่องจากข้อมูลมีโอกาสเป็นสมาชิกเพียงกลุ่มใดกลุ่มหนึ่งเท่านั้น การจัดกลุ่มแบบฟัซซี่นั้น สมาชิกของกลุ่มมีโอกาส หรือค่าการเป็นสมาชิกของ

ข้อมูลระดับต่าง ๆ ในทุกกลุ่ม สำหรับการแบ่งกลุ่มแบบฟัซซี่ (Fuzzy Clustering) ขั้นตอนการทำงานของฟัซซี่ซีมีน ประกอบด้วย

1. กำหนดกลุ่มข้อมูลที่ต้องการจัดกลุ่ม เพื่อกำหนดค่าเพื่อเป็นเงื่อนไขในการให้ข้อมูล หยุดการจัดกลุ่ม (ε) กำหนดค่าฟัซซี่พารามิเตอร์ (m) ซึ่งต้องมากกว่าหนึ่ง และ กำหนดจุดศูนย์กลางเริ่มต้นของข้อมูล

2. คำนวณค่าการเป็นสมาชิกของข้อมูลต่อกลุ่มข้อมูลต่างๆ

3. คำนวณจุดศูนย์กลางกลุ่มข้อมูลใหม่และตรวจสอบเงื่อนไขโดยตรวจสอบค่าการเป็นสมาชิกใหม่ลบค่าการเป็นสมาชิกก่อนหน้า

4. ถ้าเงื่อนไขเป็นจริงคำนวณค่าการเป็นสมาชิกและสมการเป้าหมาย (Objective Function) ถ้าเงื่อนไข เป็นเท็จ คำนวณค่าการเป็นสมาชิกจากจุดศูนย์กลางล่าสุด (วนรอบ) การคำนวณ Objective Function สามารถคำนวณจาก ดังสมการ (2)

$$J=\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu ij)^{m}\,d^{2}(X_{j},Z_{i}) \qquad (2)$$

โดย *J* แทน Objective Function ของขั้นตอนวิธีฟัซซี่ซีมีน

กำหนดให้เซตของข้อมูล $X=\{X_1,X_2,...X_n\}$

n แทน จำนวนข้อมูล

c แทน จำนวนกลุ่มข้อมูล

m แทน ฟัซซี่พารามิเตอร์ที่ต้องมีค่ามากกว่า 1

μij แทน ค่าการเป็นสมาชิกของข้อมูลที่ *J* ในกลุ่มที่ *i*

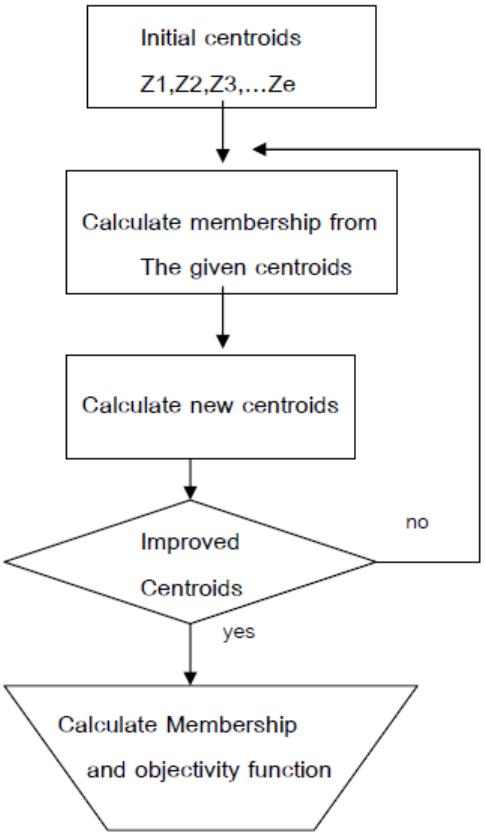
$d^{2}(X_{j},Z_{i})$ แทนระยะทางยกกำลังสองระหว่างข้อมูล *x* ที่ *j* และจุดศูนย์กลางของข้อมูล *z* กลุ่มที่ *i* โดย ดังสมการ (3)

$$Z_{i}=\frac{\sum_{j=1}^{n}(\mu ij)^{m}X_{j}}{\sum(\mu ij)^{m}} \qquad (3)$$

การหาค่าการเป็นสมาชิก μij แสดงได้ดังสมการ (4)

$$\mu ij=\frac{[1/d^{2}(X_{j}-Z_{i})]^{1/(m-1)}}{\sum_{i=1}^{c}[1/d^{2}(X_{j}-Z_{i})]} \qquad (4)$$

รายละเอียดการทำงานของฟัซซี่มีนมีการทำงานดังนี้



รูปที่ 3 การทำงานของฟัซซีซีมีน

สำหรับการวัดระยะทางระหว่างข้อมูลและจุดศูนย์กลางของข้อมูลนั้น แบบยูคลิเดียน (Euclidean Distance) ดังสมการ (5)

$$ED_{ji} = \sqrt{(X_j - Z_i)(X_j - Z_i)^T} \quad (5)$$

โดย ED_{ji} i แทนระยะทางแบบยูคลิเดียนระหว่างข้อมูล X ที่ j และจุดศูนย์กลางข้อมูล Z กลุ่มที่ i และ T แทน Transpose Matrix

สำหรับการวัดระยะทางแบบมหาลาโนบิส (Mahalanobis Distance) นั้นเหมาะกับข้อมูล ที่มีความสัมพันธ์ต่อกัน สามารถหาค่าได้จากสมการ (6)

$$MD_{ji} = \sqrt{(X_j - Z_i)A^{-1}(X_j - Z_i)^T} \quad (6)$$

โดย MD_{ji} แทนระยะทางแบบมหาลาโนบิสระหว่างข้อมูล X ตัวที่ j และจุดศูนย์กลางข้อมูล Z กลุ่มที่ i

A คือ Variance-Covariance Matrix คำนวณจากสมการ (7)

$$A = \frac{\sum_{j=1}^{n} (X_j - Z_i)^T (X_j - Z_i)}{n-1} \quad (7)$$

Esfahani และ Alhan [25] ได้การแก้ไขปัญหาการเริ่มต้นของข้อมูล โดยการนำข้อมูลคุณสมบัติของข้อมูลและข้อมูลผู้ใช้มาทำการจัดกลุ่มด้วยวิธีการ C-means ก่อนจะเข้าสู่กระบวนการกลั่นกรองแบบพึ่งพาผู้ใช้ร่วม Li และ Kim (1) การทำการแก้ไขปัญหาการเริ่มต้นของข้อมูล โดยการจัดกลุ่มข้อมูลคุณสมบัติของภาพยนตร์ ด้วยวิธีการฟัซซี่ ประกอบด้วยประเภทของภาพยนตร์ นักแสดงชาย หญิง และผู้กำกับ เมื่อทำการจัดกลุ่มเรียบร้อยแล้วทำการคำนวณค่าความใกล้เคียงต่อไป

2.3 ประเภทของระบบแนะนำข้อมูล

การประมวลผลระบบแนะนำข้อมูล เป็นที่น่าสนใจของนักวิจัยอยู่ เนื่องด้วยการวิจัยในสาขานี้สามารถก่อให้เกิดแอปพลิเคชั่นมากมายที่ช่วยลดปัญหาข้อมูลจำนวนมาก (Information Overload) ให้กับผู้ใช้ได้และยังเป็นการแนะนำข้อมูลที่มีความเป็นเฉพาะบุคคลได้ด้วยระบบแนะนำเข้ามามีบทบาทเมื่อกลางปี ค.ศ.1990 [27] โดยถูกแบ่งเป็นประเภทตามวิธีการสร้างการแนะนำ [1] [28] ได้แก่

2.3.1 การคัดกรองข้อมูลแบบอิงเนื้อหา (Content-based Filtering : CBF) เป็นเทคนิคที่มีพื้นฐานจากข้อมูลที่ได้จากเนื้อหาของชิ้นข้อมูลนั้น ๆ ซึ่งเทคนิคนี้จะให้ความสนใจกับคุณลักษณะของข้อมูลเป็นสำคัญ [29] [30] เช่น คุณลักษณะพื้นฐาน (Feature) เพื่อค้นหาชิ้นข้อมูลที่ผู้ใช้สนใจ ซึ่งชิ้นของข้อมูลที่มีลักษณะตรงตามความชอบของผู้ใช้ จึงจะถูกแนะนำให้แก่ผู้ใช้ ซึ่งจะต้องใช้ขั้นตอนวิธีการที่เกิดจากการเรียนรู้ของเครื่อง โดยจะสนใจว่าลักษณะชิ้นข้อมูลนั้นตรงตามความชอบของผู้ใช้หรือไม่ ซึ่งถ้าใช่ก็จะทำการแนะนำชิ้นข้อมูลดังกล่าว แต่ถ้าไม่ใช่ก็จะไม่ทำการแนะนำให้แก่ผู้ใช้เลย ซึ่งจากลักษณะต่าง ๆ ที่เป็นรายละเอียดของชิ้นข้อมูลนั้น ๆ สามารถบ่งชี้ได้ว่าผู้ที่มีความชอบในลักษณะของภาพยนตร์แบบนี้ ก็อาจจะมีความชอบในภาพยนตร์ที่มีลักษณะที่คล้ายคลึงกัน ขั้นตอนวิธีการที่ใช้ในการวิเคราะห์ข้อมูลของภาพยนตร์นั้น โดยวิธีการจะเป็นการคำนวณหาค่าความคล้ายคลึงระหว่างข้อมูลภาพยนตร์กับโปรไฟล์ของผู้ใช้ เป็นการนำฟังก์ชั่นที่ใช้ใน

การจัดกลุ่มข้อมูลผู้ใช้หรือข้อมูลภาพยนตร์ วิธีการประมวลผล เช่น วิธีการแบบ KNN การจัดกลุ่มข้อมูล นาอีฟเบย์ [24] เป็นต้น

2.3.1.1 การกลั่นกรองข้อมูลผู้ใช้

ลักษณะของข้อมูลผู้ใช้ เช่น อายุ เพศ อาชีพ นั้น ถือว่าเป็นข้อมูลสำคัญ เนื่องจากในปัญหาของระบบแนะนำข้อมูล มีปัญหาเริ่มต้น คือ ไม่มีข้อมูลของผู้ใช้หรือผู้ใช้ใหม่ (New User) ซึ่งจะส่งผลให้ระบบไม่สามารถแนะนำข้อมูลผู้ใช้ หรือลูกค้าที่มาใหม่ได้ ดังนั้น ระบบแนะนำข้อมูลจึงได้มีการนำข้อมูลผู้ใช้มาพิจารณากลั่นกรองในการแก้ไขปัญหาผู้ใช้ใหม่

2.3.1.2 การกลั่นกรองข้อมูลลักษณะของภาพยนตร์

ลักษณะข้อมูลของภาพยนตร์ เช่น ข้อมูลประเภท ผู้กำกับ นักแสดงชาย นักแสดงหญิง ถือว่าเป็นข้อมูลที่สำคัญอีกข้อมูลเนื่องจากในปัญหาของระบบแนะนำข้อมูล มีปัญหาเริ่มต้น คือ ไม่มีข้อมูลของชิ้นข้อมูลใหม่หรือภาพยนตร์ใหม่ (New Item) ซึ่งจะส่งผลให้ระบบไม่สามารถแนะนำข้อมูลภาพยนตร์ใหม่ได้ ดังนั้น ระบบแนะนำข้อมูลจึงได้มีการนำลักษณะข้อมูลของภาพยนตร์มาพิจารณากลั่นกรองในการแก้ไขปัญหาข้อมูลใหม่หรือภาพยนตร์ใหม่ มี

สรุปข้อดี ข้อเสียด้วยวิธีการ CBF ดังนี้

ข้อดี

1) ไม่ต้องอาศัยความเห็นจากการให้ระดับการให้คะแนนของผู้ใช้ เหมาะกับผู้ใช้ที่มีความชอบไม่เหมือนกระแสนิยมทั่วไปและสามารถแนะนำชิ้นข้อมูล โดยไม่ต้องอาศัยความเห็นจากการใช้งานของผู้ใช้ระบบอื่นที่อยู่ในระบบ

2) ไม่พบข้อเสียเกี่ยวกับ First-rater Problem คือ ถ้าเราเป็นผู้ใช้คนแรกในระบบ เราจะไม่มีเพื่อน เมื่อไม่มีเพื่อนก็จะไม่มีผลของการแนะนำ และค่าคะแนนเบาบาง ปัญหาคือถ้ามีการใส่ค่าความชอบของแต่ละผู้ใช้ที่มีต่อแต่ละสินค้าในจำนวนที่น้อยมาก จะทำให้จำนวนการให้ค่าความชอบในสินค้าเดียวกัน (Co-rated Items) มีจำนวนน้อยตามไปด้วย ทำให้เพื่อนที่ได้มีคุณภาพต่ำ

ข้อเสีย

1) ในการแนะนำด้วยเทคนิคการคัดกรองข้อมูลแบบอิงเนื้อหานั้น ระบบจะไม่สามารถแนะนำสิ่งที่แตกต่างไปจากสิ่งที่ผู้ใช้เคยใช้งานหรือมีประสบการณ์กับชิ้นข้อมูลนั้นมาก่อน ทำให้ไม่สามารถแนะนำสิ่งที่แตกต่างไปจากสิ่งที่ผู้ใช้งานมีความชอบได้ ทั้งที่ชิ้นข้อมูลดังกล่าวอาจเป็นชิ้นข้อมูลที่มีความนิยมกับผู้ใช้คนอื่น ๆ ด้วย

2) การหาความสัมพันธ์ระหว่างข้อมูลผู้ใช้และรายการสินค้า สิ่งที่จำเป็นอย่างยิ่งคือการที่ต้องดึงคุณสมบัติต่าง ๆ ของสินค้าออกมาใช้ แต่การดึงคุณสมบัติที่เหมาะสมเป็นสิ่งที่อยาก

และเป็นปัญหาของนักวิจัยอยู่ขณะนี้ เพราะถ้าหาคุณสมบัติออกมาไม่ดี ค่าความสัมพันธ์ที่ได้ก็จะไม่ถูกต้อง ซึ่งจะทำให้ผลลัพธ์ของการแนะนำไม่ถูกต้องตามไปด้วย

2.3.2 การคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering : CF) ผู้ใช้จะได้รับการแนะนำชิ้นข้อมูลหรือสินค้าที่ผู้ใช้ท่านอื่นที่มีรสนิยมคล้ายกันเคยมีประสบการณ์ที่ดีกับชิ้นข้อมูลนั้นในอดีต เทคนิคที่เคยใช้ในงานวิจัยได้แก่ ค่าความใกล้เคียง การจัดกลุ่ม ทฤษฎีกราฟ การวิเคราะห์การถดถอยเชิงเส้น เป็นต้น

2.3.2.1 Item Based เป็นวิธีการหาเพื่อนบ้านที่มีลักษณะการให้คะแนน ค่าคะแนนใกล้เคียง (Similarity) กับผู้ใช้มาเพื่อใช้ในการทำนาย ค่าคะแนนให้กับชิ้นข้อมูลที่ผู้ใช้ยังไม่ได้ให้ค่าคะแนน โดยการนำข้อมูลการให้คะแนน ค่าคะแนนในอดีตมาเปรียบเทียบกับกลุ่มของผู้ใช้ที่มีลักษณะการให้คะแนน ค่าคะแนน ใกล้เคียงกันเพื่อที่จะทำนายและนำเสนอชิ้นข้อมูลที่คาดว่าผู้ใช้จะสนใจมาแนะนำให้กับผู้ใช้ตามความสนใจของผู้ใช้ ซึ่งมีอัลกอริทึมที่เกี่ยวข้อง ดังนี้

1. Adjusted Cosine-based Similarity

Adjusted Cosine-based Similarity เป็นการคำนวณหาค่าความคล้ายคลึงโดยมีแนวคิดว่าผู้ใช้แต่ละคนจะให้คะแนนต่างกันบางคนให้คะแนนสูงหมดทุกรายการ บางคนให้คะแนนต่ำ หมดทุกรายการกล่าวคือ มีผู้ใช้ที่ให้คะแนนกลางหรือสูงมากเต็มไปหมด ฉะนั้นค่าที่ได้ควรจะทำให้เป็นมาตรฐานก่อนโดยการปรับค่าเฉลี่ย ของผู้ใช้เองโดยการคำนวณนั้นจะใช้พื้นฐานของการวัดแบบโคไซน์โดยคำนวณได้จากสมการ (8)

$$sim(t,c) = \frac{\Sigma u \in U(R_{u,t} - \bar{R}_u)(R_{u,c} - \bar{R}_u)}{\sqrt{\sum u \in U(R_{u,t} - \bar{R}_u)^2}\sqrt{\sum u \in U(R_{u,c} - \bar{R}_u)^2}} \quad (8)$$

เมื่อ *sim (t,c)* แทนค่าความคล้ายคลึงระหว่างชิ้นข้อมูล *t* กับ *c*

$R_{u,t}$ และ $R_{u,c}$ แทนระดับการให้คะแนนที่ผู้ใช้ *u* มีต่อชิ้นข้อมูล *t* และผู้ใช้ *u* ต่อชิ้นข้อมูล *c*

t แทนชิ้นข้อมูลเป้าหมาย

u แทนชิ้นข้อมูลเปรียบเทียบ

$\bar{R}_u$ แทนค่าเฉลี่ยระดับการให้คะแนนของผู้ใช้งาน *u*

2. Cosine-based Similarity

Cosine-based Similarity เป็นวิธีที่คำนวณหาค่าความคล้ายคลึง โดยชิ้นข้อมูล 2 ชิ้นถูกมองเหมือนเวกเตอร์ 2 เส้น ใน มิติของพื้นที่ของผู้ใช้งาน ความคล้ายคลึงระหว่างชิ้นข้อมูลทั้งสอง วัดได้โดยคำนวณค่าโคไซน์ของมุมระหว่างเวกเตอร์ทั้ง 2 เส้น ซึ่งสามารถหาได้จากการคำนวณตามสมการ (9)

$$sim(t,c) = \frac{\sum u \in UR_{u,t} \times \bar{R}_u, c)}{\sqrt{\sum u \in UR_{u,t}^2}\sqrt{\sum u \in UR_{u,c}^2}} \quad (9)$$

เมื่อ $sim(t,c)$ แทนค่าความคล้ายคลึงระหว่างชิ้นข้อมูล t กับ c

$R_{u,t}$ และ $R_{u,c}$ แทนคะแนนจัดอันดับที่ผู้ใช้ u มีต่อชิ้นข้อมูล t และผู้ใช้ u ต่อชิ้นข้อมูล c

t แทนชิ้นข้อมูลเป้าหมาย

u แทนชิ้นข้อมูลเปรียบเทียบ

U แทน ชุดของผู้ใช้ที่ให้คะแนนทั้งรายการ

4. Correlation Based Similarity

การคำนวณหาค่าความคล้ายคลึงด้วยวิธี Correlation Based Similarity การคำนวณด้วยวิธีนี้เป็นการหาค่าความคล้ายคลึงระหว่าง 2 ชิ้นข้อมูลแทนด้วย u และ v นั้น วัดด้วยการคำนวณค่าความสัมพันธ์ของ Pearson หรือเรียกว่า $corr_{i,t}$ การคำนวณค่าความสัมพันธ์ให้ถูกต้องนั้น เราต้องแยกกรณีแรกของโคเรต (กรณีเมื่อผู้ใช้ให้คะแนนจัดอันดับทั้ง u และ v) ซึ่งสามารถคำนวณได้จากสมการ (10)

$$sim(u,v) = \frac{\sum_{i \in I}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_{i \in I}(r_{v,i} - \bar{r}_v)^2}} \quad (10)$$

เมื่อ $sim(u,v)$ แทนค่าความคล้ายคลึงระหว่างชิ้นข้อมูล u กับ v

$r_{u,i}$ และ $r_{v,i}$ แทนคะแนนชิ้นข้อมูล u มีต่อชิ้นข้อมูล i และคะแนนชิ้นข้อมูล v ต่อชิ้นข้อมูล i

$\bar{r}_u$ และ $\bar{r}_v$ แทนค่าเฉลี่ยของชิ้นข้อมูล u และ v

$\sum i \in I$ แทนผลรวมของชิ้นข้อมูลทั้งหมด

4. Jaccard Similarity หรือที่เรียกว่าค่าสัมประสิทธิ์ความคล้ายคลึงกัน Jaccard โดย Paul Jaccard เป็นผู้เริ่มในการนำมาใช้ เป็นสถิติที่ใช้สำหรับการเปรียบเทียบความคล้ายคลึงกันและความหลากหลายของชุดตัวอย่างค่าสัมประสิทธิ์ Jaccard มาตรการความคล้ายคลึงกันระหว่างชุดตัวอย่าง และถูกกำหนดให้เป็นขนาดตัดแบ่งตามขนาดของชุดตัวอย่าง

สูตรสมการ

แบบไบนารี่เวคเตอร์ (Binary Term Vector)

$$J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{11}$$

แบบให้น้ำหนักเวคเตอร์ (Weighted Term Vector)

$$J(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n} B_i^2 - \sum_{i=1}^{n} A_i B_i} \tag{12}$$

เมื่อ $0 \le J(A,B) \le 1$

$J(A,B)$ แทนค่าความคล้ายคลึงระหว่างชิ้นข้อมูล *A* กับ *B*

A แทนชิ้นข้อมูลเป้าหมาย

B แทนชิ้นข้อมูลเปรียบเทียบ

n แทนชิ้นข้อมูลทั้งหมด

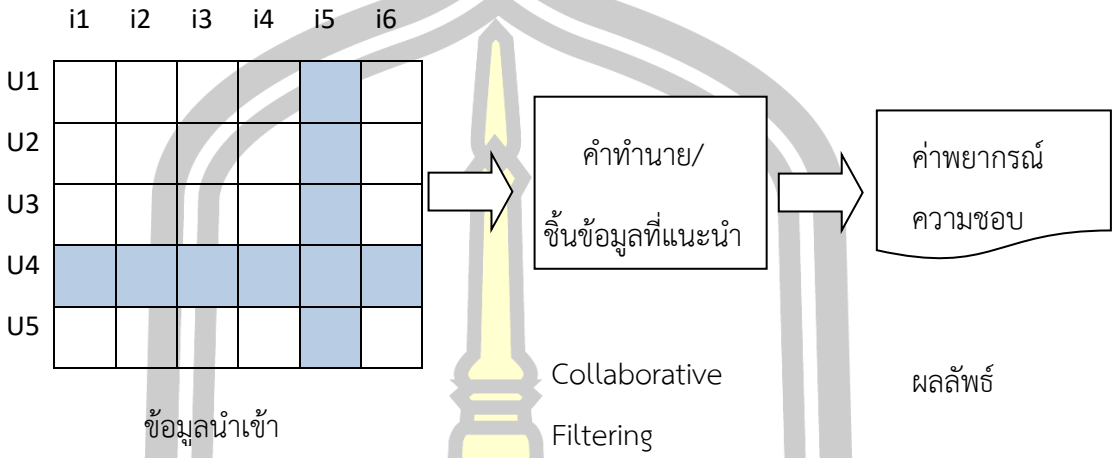
นอกจากนี้ Jaccard ยังสามารถหาระยะทางได้ (Jaccard Distance) โดยการนำ 1 มาลบกับค่าสัมประสิทธิ์ Jaccard มีสูตรคำนวณ ดังนี้

$$d_j(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{13}$$

Jaccard distance เป็นอัตราส่วนของขนาดของความแตกต่าง $A \Delta B = (A \cup B) - (A \cap B)$

ทั้งสี่วิธีนี้เป็นวิธีการหาค่าความคล้ายคลึงของทั้งหมด ต่างกันเพียงแค่สิ่งที่นำมาคำนวณในสมการ เราหาค่าความสัมพันธ์เพื่อที่จะรู้ได้ว่าเพื่อนบ้านมีความคล้ายคลึงกับผู้ใช้มาก

เพียงใดจากนั้นเลือก กลุ่มผู้ใช้มาจำนวนหนึ่งเพื่อทำการทำนายชิ้นข้อมูลที่ผู้ใช้ยังไม่ได้ทำการให้ค่าคะแนน



รูปที่ 4 กระบวนการทำงานของการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม

จากรูปที่ 4 แสดงกระบวนการทำงานของการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม มี 3 ขั้นตอน คือส่วนอินพุท ส่วนกระบวนการอัลกอริทึม และส่วนผลลัพธ์ รายละเอียดดังนี้

ขั้นตอนที่ 1 อินพุท เป็นข้อมูลที่ใช้ในระบบอยู่ในรูปของเมตริกซ์ของคะแนน ค่าคะแนน ของผู้ใช้ส่งเข้ามาในระบบดังตัวอย่าง ในตาราง 1

ตาราง 1 เมตริกซ์ของผู้ใช้และภาพยนตร์

User/Item	Movie1	Movie2	Movie3	Movie4
U1	2		5	4
U2		4		1
U3	3	4	4	
U4		5		
U5	4			

จากตาราง 1 เป็นตัวอย่างของการให้คะแนนความชอบของภาพยนตร์ที่เกิดจากผู้ใช้ 5 คน ต่อภาพยนตร์ 4 เรื่อง ซึ่งแต่ละช่องคือค่าคะแนน (1-5) ที่ผู้ใช้ให้แก่ภาพยนตร์เรื่องนั้น ช่องที่ไม่มีคะแนน ค่าคะแนน คือ ภาพยนตร์ที่ผู้ใช้ยังไม่ได้ทำการให้ ค่าคะแนน เมื่อพิจารณาหาความใกล้เคียง

แบบ Item Based ของ User1จะเห็นว่า ผู้ใช้คนนี้มีการดูภาพยนตร์ เรื่องที่ 1 2 และ 4 ซึ่งมีความใกล้เคียงกับการดูภาพยนตร์ของ User3 ที่มีการดูภาพยนตร์เรื่องที่ 1 2 และ 4 ดังนั้นเราสามารถสรุปได้ U1= $S_{m1,m3}$ (2,5) ใกล้เคียงกับ U3= $S_{m1,m3}$ (3,4) ซึ่งมีการให้ค่าคะแนนที่แตกต่างกัน

ขั้นตอนที่ 2 เข้าสู่กระบวนการประมวลผลอัลกอริทึมเป็นขั้นตอนที่สำคัญที่สุดของเทคนิคการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม โดยจะทำเมตริกซ์ของค่าคะแนนและเมทตริกซ์ของค่าความใกล้เคียงเพื่อมาทำการหาค่าคะแนน ดังสมการ (14)

$$pred(u,i) = \frac{\sum_{j \in ratedItems(u)} itemSim(i,j).r_{uj}}{\sum_{j \in ratedItems(u)} |itemSim(i,j)|} \qquad (14)$$

เมื่อ $temSim(i,j)$ คือ ค่าความใกล้เคียงของชิ้นข้อมูล i กับ j

r_{uj} คือ ค่าคะแนนของผู้ใช้ u กับชิ้นข้อมูล j

ขั้นตอนที่ 3 เป็นขั้นตอนการแสดงผลที่ได้จากการทำนายแนะนำข้อมูลที่คาดว่าผู้ใช้มีต่อชิ้นข้อมูลเป้าหมายโดยจะพิจารณาข้อมูลภาพยนตร์มีค่าคะแนนสูงสุดไปแนะนำให้กับผู้ใช้คนอื่นได้

2) User based เป็นการพัฒนารูปแบบการกรองข้อมูล ซึ่งจะให้เทคนิคทางดาต้าไมนิ่งอีกวิธีหนึ่ง คือมีข้อมูลที่ใช้ในการสร้างโมเดลและข้อมูลที่ใช้ในการทดสอบโมเดล อัลกอริทึมที่นิยมนำมาใช้ เช่น Bayesian Collaborative Filtering หรือ Clustering Collaborative Filtering หรือ Latent Sematic Collaborative Filtering หรือ Collaborative Filtering using dimensionality reduction ซึ่งวิธีการนี้มีข้อดีคือสามารถแก้ไขปัญหาข้อมูลเบาบาง และข้อมูลจำนวนมากได้ แต่ข้อเสียคือ สูญเสียข้อมูลที่เป็นประโยชน์ในการลดมิติ

ตาราง 2 เมตริกซ์ของผู้ใช้และภาพยนตร์

User/Item	Movie1	Movie2	Movie3	Movie4
U1	2		5	4
U2		4		1
U3	3	4	4	
U4		5		
U5	4			

จากตาราง 2 เป็นตัวอย่างของการให้คะแนนความชอบของภาพยนตร์ที่เกิดจากผู้ใช้ 5 คน ต่อภาพยนตร์ 4 เรื่อง ซึ่งแต่ละช่องคือค่าคะแนน (1-5) ที่ผู้ใช้ให้แก่ภาพยนตร์เรื่องนั้น ช่องที่ไม่มีคะแนน ค่าคะแนน คือ ภาพยนตร์ที่ผู้ใช้ยังไม่ได้ทำการให้ ค่าคะแนน เมื่อพิจารณาหาความใกล้เคียงแบบ User Based ของ User1 และ User3 มีข้อมูลการดูภาพยนตร์ เรื่องที่ 1 และเรื่องที่ 3 ดังนั้น เราสามารถสรุปได้ Movie1= $S_{u1,u3}$ (2,3) ใกล้เคียงกับ Movie3= $S_{u1,u3}$ (5,4) ซึ่งมีการให้ค่าคะแนนที่แตกต่างกัน หลังจากคำนวณค่าความใกล้เคียงแล้ว จะทำการหาค่าคะแนนน้ำหนักที่มีความใกล้เคียงกันผู้ใช้แต่ละคนกับผู้ใช้เป้าหมายที่กำหนด ดังสมการ (15)

$$P(u,i) = \overline{r}_u + \frac{\sum_{u_k \in S(u)} Sim(u,u_k).(r_{u_k,i} - \overline{r}_{u_k})}{\sum_{u_k \in S(u)} Sim(u,u_k)} \qquad (15)$$

เมื่อ $P(u,i)$ แทน ค่าน้ำหนักรวมของผู้ใช้ u ต่อชิ้นข้อมูล i

$\overline{r}_u$ แทนค่าเฉลี่ยของผู้ใช้ u

$Sim(u,u_k)$ แทนค่าความใกล้เคียงของคู่ผู้ใช้ u และ u_k

$r_{u_k,i}$ แทนค่าคะแนนของผู้ใช้ u_k ที่มีต่อชิ้นข้อมูล i

$\overline{r}_{u_k}$ แทนค่าคะแนนเฉลี่ยของผู้ใช้ u_k

$S(u)$ แทนค่าความใกล้เคียงของผู้ใช้ u

สรุปข้อดี ข้อเสียด้วยวิธีการ CF ดังนี้

เนื่องจากอัลกอริทึมการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม อาศัยโคเรตในการค้นหาผู้ใช้ที่มีลักษณะใกล้เคียงกัน จึงเป็นสาเหตุสำคัญให้เกิดปัญหาสำคัญ หลายประการด้วยกัน ดังต่อไปนี้

ข้อดี

การแนะนำด้วยเทคนิคการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมนั้น ระบบจะสามารถแนะนำสิ่งที่แตกต่างไปจากสิ่งที่ผู้ใช้เคยใช้งานหรือมีประสบการณ์กับชิ้นข้อมูลนั้นมาก่อน ทำให้สามารถแนะนำสิ่งที่แตกต่างไปจากสิ่งที่ผู้ใช้งานมีความชอบได้ ทั้งที่ชิ้นข้อมูลดังกล่าวอาจเป็นชิ้นข้อมูลที่มีความนิยมกับผู้ใช้คนอื่น ๆ ด้วย

ข้อเสีย

1) ปัญหาขนาดของข้อมูล (Scalability Problem)

ปัจจุบันการขยายตัวของข้อมูล เป็นความท้าทายสำหรับกระบวนการออกแบบ การพยากรณ์ ค่าคะแนน สามารถเกิดขึ้นได้จาก ปริมาณของผู้ใช้งานที่เพิ่มขึ้นตามเวลา ในการปรับปรุง Item-Item Similarity สำหรับกรณีนี้ จะเป็นการคำนวณที่สิ้นเปลืองมาก เนื่องจากคุณสมบัติของข้อมูลจะมีการเปลี่ยนแปลงที่น้อยกว่าข้อมูลผู้ใช้ ดังนั้นจึงสามารถคำนวณจากค่าความใกล้เคียงของข้อมูลไว้ล่วงหน้าก่อนได้ ส่วนข้อมูลผู้ใช้นั้นต้องการการคำนวณที่มากกว่า ในกรณีที่ระบบมีขนาดใหญ่ขึ้น การแก้ปัญหานี้สามารถทำได้โดยระบบสามารถใช้วิธีการแบ่งกลุ่มผู้ใช้งาน (User Segmentation) ด้วยวิธี จัดกลุ่มก่อนเข้าสู่กระบวนการการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม เพื่อให้ได้มาซึ่งผลลัพธ์ที่จะแนะนำสู่ผู้ใช้งานได้เร็วขึ้น

2) ปัญหาชิ้นข้อมูลที่ไม่มีการให้ค่าคะแนน ไว้ (First-rater Problem)

เป็นปัญหาที่เกิดจากชิ้นข้อมูลใหม่หรือชิ้นข้อมูลที่ยังไม่มีผู้ใช้คนใดเคยให้ค่าคะแนน มาก่อน ทำให้ชิ้นข้อมูลชิ้นนั้นไม่สามารถนำมาเปรียบเทียบกับชิ้นข้อมูลใด ๆ ได้เลย ดังนั้นจึงไม่สามารถนำมาคำนวณหาค่าความพึงพอใจได้ ตัวอย่างของปัญหาของระบบแนะนำภาพยนตร์ก็คือ ภาพยนตร์เรื่องใหม่ที่ยังไม่ได้ฉายหรือยังไม่มีผู้ใช้คนใดทำการให้ ค่าคะแนน จะไม่มีโอกาสที่จะถูกหยิบออกมาแนะนำให้กับผู้ใช้คนใดได้เลย

2.3.4 การกลั่นกรองข้อมูลแบบผสมผสาน (Hybrid Filtering : HF)

การกลั่นกรองข้อมูลแบบผสมผสาน เป็นวิธีการรวมเอาการคัดกรองข้อมูลแบบอิงเนื้อหา และการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม มาผสมผสานกัน หรือการนำวิธีการแก้ไขปัญหาจากอื่นที่นำวิธีการแก้ไขปัญหาตั้งแต่สองวิธีขึ้นไป เพื่อสร้างการแนะนำชิ้นข้อมูลหรือสินค้าแก่ผู้ใช้งาน เพื่อลดข้อเสียและเสริมข้อดีของทั้งสองเทคนิคดังที่ได้กล่าวไว้ข้างต้น[27] [31] ตัวอย่าง เฟรมเวิร์คในการทำงานร่วมกัน เช่น Weighted, Switching, Cascade และ Feature Combination เป็นต้น ซึ่งแต่ละเฟรมเวิร์คเหมาะสมกับการใช้งานที่แตกต่างกัน เช่น Weighted เหมาะสมกับกรณีเมื่อต้องการให้ความสำคัญของแต่ละเทคนิคแตกต่างกัน ซึ่งทำได้ง่ายโดยปรับค่าน้ำหนัก อีกหนึ่งตัวอย่างของการทำงานร่วมกัน เช่น Switching ใช้วิธีเปลี่ยนเทคนิคที่ใช้ในการสร้างคำทำนายตามสภาวะแวดล้อม เป็นผลให้เทคนิคที่ใช้ในการสร้างคำทำนายเป็นเทคนิคที่เหมาะสมสุดสำหรับสภาวะ

แวดล้อมขณะนั้น ดังนั้นการเลือกใช้รูปแบบการทำงานร่วมกันควรพิจารณาตามสมควร ซึ่งการกลั่นกรองข้อมูลแบบผสมผสานมีหลายวิธี (2) ดังนี้

2.3.4.1 การทำงานร่วมกันแบบถ่วงน้ำหนัก (Weighted)

การทำงานร่วมกันแบบถ่วงน้ำหนัก [25] คือ เทคนิคที่สร้างคำทำนายโดยใช้ทุกตัวทำนายที่มีอยู่ในระบบ จากนั้นนำคำทำนายที่ได้ทั้งหมดจากทุกตัวทำนาย มาคำนวณเพื่อหาคำทำนายสุดท้าย คำนวณโดยเทคนิคการให้น้ำหนักแก่คำทำนาย ตัวอย่างระบบที่สร้างคำทำนายโดยเทคนิคการทำงานร่วมกันแบบถ่วงน้ำหนัก เช่น ระบบ P-Tango System ซึ่งหาคำทำนายด้วยวิธีการรวมแบบเชิงเส้น (Linear Combination) โดยค่าน้ำหนักของคำทำนายสามารถปรับตามความถูกต้องของคำทำนายเปรียบเทียบกับคำตอบที่ได้รับจากผู้ใช้งาน ระบบ Pazzani เป็นอีกระบบที่สร้างคำทำนายโดยใช้การทำนายร่วมกันแบบถ่วงน้ำหนัก โดยพิจารณาคำทำนายซึ่งอยู่ในลักษณะประเภท (Nominal) ไม่ใช่ตัวเลข (Numeric) ดังนั้นการหาคำทำนายสุดท้ายจึงใช้วิธีคะแนนเสียงส่วนใหญ่ (Majority Vote) ข้อดีของการทำงานร่วมกันแบบถ่วงน้ำหนัก คือ ได้ใช้ความสามารถของแต่ละตัวทำนายอย่างเต็มที่ และการปรับเปลี่ยนความสำคัญของแต่ละตัวทำนายทำได้ง่ายและตรงไปตรงมาโดยทาการปรับค่าน้ำหนักนั่นเอง

2.3.4.2 การทำงานแบบสลับ (Switching)

การทำงานการแนะนำข้อมูลที่มีความซับซ้อนเข้าสู่กระบวนการให้คำแนะนำตั้งแต่การเปลี่ยนเกณฑ์ หรือการแนะนำชิ้นวัตถุมีความถูกต้องน้อยกว่าอีกวิธีหนึ่ง จะต้องพิจารณาสลับเป็นอีกวิธีที่มีความถูกต้องมากที่สุด ซึ่งวิธีการนี้เป็นวิธีการที่สามารถมีความไวต่อจุดแข็งและจุดอ่อนของการแนะนำข้อมูล วิธีการแบบสลับจะเป็นการนำข้อดีของวิธีการคัดกรองแบบอิงเนื้อหา และวิธีการคัดกรองแบบพึ่งพาผู้ใช้ร่วมมาใช้ เพื่อที่จะลดข้อเสียของอีกวิธีหนึ่ง ทำให้สามารถแก้ไขปัญหาของข้อเสียของทั้งสองวิธีได้ [27]

2.3.4.3 การผสม (Mixed)

การผสมข้อมูลที่มีจำนวนมาก โดยนำค่าน้ำหนักมารวมกันแล้วหาค่าเฉลี่ย ซึ่งไฮบริดผสมหลีกเลี่ยงปัญหาเริ่มต้น คือ รายการใหม่ ซึ่งยังไม่ได้จัดอันดับคะแนนจากข้อมูลผู้ใช้และผู้ใช้ใหม่ ซึ่งผู้ใช้ยังไม่มีพฤติกรรมต่อสินค้า ทำให้ไม่มีค่าคะแนนมาจัด

2.3.4.4 รวมคุณลักษณะ (Feature Combination) ระบบที่เป็นไปตามวิธีการรวมกันคุณลักษณะของ Content based และ Collaborative ที่มีคุณลักษณะใกล้เคียงกันมีความสัมพันธ์กัน ซึ่งวิธีการนี้พบว่าคุณลักษณะข้อมูลบางอย่างสามารถช่วยเสริมหรือปรับปรุงอีกวิธีการหนึ่งให้ระบบแนะนำข้อมูลมีความถูกต้องมาก ไฮบริดรวมคุณลักษณะที่รวมกันจะช่วยให้ระบบการทำงานร่วมกันพิจารณาข้อมูลโดยไม่ต้องอาศัยเฉพาะตัวนั้น ดังนั้น จึงช่วยลดจุดด้อยของระบบ

ที่อาศัยจำนวนผู้ใช้ที่ได้รับการจัดอันดับรายการหรือค่าคะแนน ซึ่งวิธีการนี้จะช่วยให้ระบบมีข้อมูลเกี่ยวกับความคล้ายคลึงกันเพิ่มมากขึ้น

2.3.4.5 Cascade แนวคิดวิธีการนี้จะให้ความสำคัญของค่าน้ำหนักที่มีค่าน้อยหรือมากเท่ากัน โดยค่าคะแนนน้อยอาจจะถูกจัดอันดับคะแนนเพิ่มขึ้นหรือจัดอันดับค่าน้ำหนักใหม่ ตามค่าน้ำหนักที่มีค่ามากกว่าก็ได้ วิธีการนี้เป็นการให้ความสำคัญของค่าน้ำหนักที่น้อย ซึ่งปกติระบบแนะนำข้อมูลจะให้ความสนใจกับค่าน้ำหนักที่มีค่ามาก เพื่อนำไปใช้ในระบบแนะนำข้อมูล และค่าที่มีน้ำหนักน้อยจะถูกตัดทิ้ง ไม่ให้ความสำคัญ แต่วิธีการนี้ ยังให้ความสำคัญกับข้อมูลทีมีค่าน้อย ซึ่งสามารถที่จะปรับค่าจากความสัมพันธ์ของข้อมูลที่มีค่ามากมาช่วยในการเพิ่มค่าน้ำหนักได้

2.3.4.6 คุณสมบัติเสริม (Feature Augmentation) วิธีการนี้อาศัยข้อมูลที่เป็นค่าคะแนนของรายการหรือจำแนกประเภทข้อมูล และเป็นการรวมข้อมูลในการประมวลผลสำหรับแนะนำข้อมูลต่อไป เช่น ระบบแนะนำหนังสือ ที่ใช้ประมวลผลด้วยวิธีการ Content based ที่มีการนำข้อมูลชื่อผู้แต่ง และชื่อเรื่องของหนังสือ ด้วยวิธีการนาอีฟเบย์ แต่ยังได้นำวิธีการแบบ Collaborative มาใช้ร่วมในระบบแนะนำข้อมูลได้อย่างมีประสิทธิภาพ วิธีการนี้เป็นที่น่าสนใจ เพราะว่าเป็นวิธีการที่ช่วยปรับปรุงประสิทธิภาพการทำงานของระบบหลัก

2.3.4.7 Meta-level ไฮบริดเมตาระดับรูปแบบทั้งหมดจะเป็นการนำเข้า โดยข้อมูลที่มีความหนาแน่นจะเหมาะกว่าข้อมูลที่เป็นคะแนนดิบที่ยังไม่ผ่านกระบวนการอะไร และสามารถที่จะนำไปประยุกต์ใช้กับวิธีการแบบ Content based และ Collaborative ได้

สรุป งานวิจัยนี้ ผู้วิจัยจะนำวิธีการแบบการทำงานร่วมกันแบบการรวมคุณลักษณะใช้ เนื่องจากในกระบวนการทำงานระบบแนะนำข้อมูลมีกระบวนการภายในที่แตกต่างกัน มีข้อมูลที่มีความหลากหลาย เป็นวิธีที่ช่วยเสริมจุดด้อยของอีกวิธีหนึ่ง จึงมีความเหมาะสมที่จะนำวิธีการนี้ไปใช้เพื่อให้ระบบแนะนำข้อมูลมีประสิทธิภาพต่อไป

2.4 การวัดประสิทธิภาพ

2.4.1 Mean Absolute Error (MAE) เป็นวิธีการหาค่าเฉลี่ยของความแตกต่างสมบูรณ์ระหว่างค่าพยากรณ์และค่าจริง ซึ่งหากผลการประเมินมีค่าน้อย แสดงว่าค่าที่พยากรณ์ได้มีความใกล้เคียงกับค่าจริง ดังสมการ (16)

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | p_{u,i} - r_{u,i} | \qquad (16)$$

เมื่อ N = จำนวนภาพยนตร์ที่แนะนำ

$p_{u,i}$ = ค่าพยากรณ์

$r_{u,i}$ = ค่าคะแนนจริง

2.4.2 ค่าความแม่นยำ (Pecision) คือ เป็นการวัดความแม่นยำในการแนะนำข้อมูลภาพยนตร์ โดยค่าความแม่นยำเป็นอัตราส่วนของชิ้นข้อมูลหรือภาพยนตร์ที่ผู้ใช้ให้ความสนใจและตรงกับชิ้นข้อมูลหรือภาพยนตร์ที่แนะนำ กับจำนวนการแนะนำชิ้นข้อมูลหรือภาพยนตร์ทั้งหมด [32] [33] สมการ (17)

$$\Pr ecision = \frac{|\{relevant\ item\} \cap \{top-N\ item\}|}{top-N\ item} \quad (17)$$

เมื่อ $relevant\ item$ คือ ชิ้นข้อมูลที่ผู้ใช้ให้ความสนใจ

$top-N\ item$ คือ ชิ้นข้อมูลที่แนะนำ

N คือ จำนวนชิ้นข้อมูลที่แนะนำ

2.5 งานวิจัยที่เกี่ยวข้อง

2.5.1 การวิเคราะห์ข้อมูล

มีงานวิจัยหลายเรื่อง [13] [32] [33] [34] ได้นำเสนอระบบแนะนำภาพยนตร์โดยข้อมูลนำเข้านักวิจัยส่วนใหญ่ใช้ข้อมูลค่าคะแนนในการนำข้อมูล นอกจากนี้มีข้อมูลผู้ใช้ประกอบด้วย อายุ เพศ อาชีพ ข้อมูลรายการภาพยนตร์ประกอบด้วยประเภท ผู้กำกับ นักแสดง ส่วนข้อมูลอื่น ๆ ประกอบด้วยคีย์เวิร์ด แท็ก เวลา เพื่อน และเงื่อนไขอื่นๆ ที่เกี่ยวข้อง เพื่อใช้ในการแนะนำข้อมูล ดังนี้

ตาราง 3 แสดงข้อมูลนำเข้า

ผู้แต่ง	ค่าคะแนน	ผู้ใช้ (User)			รายการ (Item)			บริบท (Context)				
		อายุ	เพศ	อาชีพ	ประเภท	ผู้กำกับ	นักแสดง	คีย์เวิร์ด	แท็ก	เวลา	ที่อยู่	อื่นๆ
Papagelis และ Plexousakis [33]	✓											
H.yildirim [13]	✓											
L.He, F.Wu [34]	✓									✓		
G.Pitsilis,S.J.Knapskog [35]	✓											

ตาราง 3 (ต่อ)

ผู้แต่ง	ค่าคะแนน	ผู้ใช้ (User)			รายการ (Item)			บริบท (Context)				
		อายุ	เพศ	ทักษะ	ประเภท	ผู้กำกับ	นักแสดง	คีย์เวิร์ด	แท็ก	เวลา	เพื่อน	เงื่อนไข
Yuan-hong และ TAN Xiao qui [32]	✓											
Wen, Zhou [15]	✓											
S.K.Tiwari,Shrivastava [36]	✓	✓	✓	✓	✓							
J.Gupta,J.Gudge[37]	✓	✓	✓	✓								
V.Codina และคณะ[38]	✓										✓	✓
V.Codina และคณะ[39]	✓										✓	✓
Ghazanfar และ Bennett [15]	✓		✓		✓	✓	✓	✓	✓			
Braunhofer และคณะ [10]	✓				✓	✓	✓	✓	✓			
Lam และคณะ [40]		✓	✓	✓								
Baltrunas และRicci [20]	✓	✓	✓									
Baltrunas และRicc i[6]	✓	✓										
Datta และคณะ[11]		✓	✓	✓								
Zheng และคณะ[22]	✓	✓	✓									

จากข้อมูลนำเข้าที่มีความแตกต่างกันมีนักวิจัยจำนวนมากที่ทำการทดลองในการแก้ไขปัญหาระบบแนะนำข้อมูล ซึ่งมีวิธีการที่มีความแตกต่างกัน และสอดคล้องกับข้อมูลนำเข้า เช่น

Papagelis และ Plexousakis [33] ได้ศึกษาวิเคราะห์คุณภาพของค่าความใกล้เคียงที่มีความสัมพันธ์ของผู้ใช้ ซึ่งข้อมูลนำเข้าประกอบด้วยค่าคะแนน (rating) โดยทำการเปรียบเทียบค่าความใกล้เคียงระหว่าง User-based และ Item-based ผลปรากฏว่า วิธีการ Item-based มีประสิทธิภาพดีกว่า Item-based ทั้งนี้ชี้ให้เห็นว่าความสัมพันธ์ของสินค้าในการหาค่าความใกล้เคียงมีความสัมพันธ์ที่ดีกว่าข้อมูลค่าความใกล้เคียงของผู้ใช้

Yildirim (3) ได้ศึกษาวิธีการหาความสัมพันธ์ด้วยกราฟระหว่างสินค้า ซึ่งข้อมูลนำเข้าค่าคะแนน โดยการเปรียบเทียบ Adjusted Cosine และ Cosine ซึ่งข้อมูลนำเข้าประกอบด้วยค่าคะแนน (rating) ผลปรากฏว่าวิธีการหาค่าความใกล้เคียงด้วย Adjusted Cosine มีประสิทธิภาพดีกว่า

He และ Wu [34] ได้ศึกษาวิธีการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม ด้วยวิธีการหาค่าความใกล้เคียงที่มีความสัมพันธ์ของผู้ใช้ และเพิ่มเทคนิคเงื่อนไขเวลาในการหาความถี่ ทำการเปรียบเทียบกับ วิธีการ User-based ซึ่งเมื่อมีการนำข้อมูลเวลาและหาความถี่ ทำให้มีประสิทธิภาพที่ดีกว่า User-based ที่ใช้เฉพาะข้อมูลค่าคะแนนเท่านั้น ทั้งนี้อาจเป็นเพราะว่าการเพิ่มเงื่อนไขเวลาและความถี่เข้ามาช่วยในการกลั่นกรองข้อมูลก่อนการประมวลผลแนะนำข้อมูลทำให้ประสิทธิภาพไม่แตกต่างกันมากนัก แต่งานวิจัยนี้ต้องใช้เวลาประมวลผลมากขึ้นและมีการหาค่าความใกล้เคียงเฉพาะ User-based เท่านั้น

Pitsilis และ Knapskog [35] ได้ศึกษาการเพิ่มความเชื่อมั่นและแก้ไขปัญหาข้อมูลเบาบางระบบแนะนำข้อมูล โดยอาศัยข้อมูลค่าคะแนนทำการคำนวณของมูลทางตรงและทางอ้อมที่มีความสัมพันธ์กันเปรียบเทียบกับวิธีการ Item-based ผลปรากฏว่า มีประสิทธิภาพดีกว่า

Yuan-hong และ TAN Xiao qui [32] ได้ศึกษาการแนะนำข้อมูลแบบเรียลไทม์ วิธีการคัดกรองแบบพึ่งพาผู้ใช้ร่วม โดยทำการลดมิติของข้อมูลด้วยการแยกค่าเชิงเดี่ยว (Singular Value Decomposition : SVD) และการจัดกลุ่มของข้อมูลด้วย K-mean ทำการเปรียบเทียบกับวิธีการ User-based และ Item-based ผลปรากฏว่ามีประสิทธิภาพดีกว่า ทั้งนี้อาจเป็นเพราะว่าการทำกระบวนการกรองข้อมูลก่อนเข้าสู่กระบวนการประมวลผลแนะนำข้อมูลด้วย SVD และ K-mean สามารถที่ทำให้ประสิทธิภาพดีขึ้น

Wen และ Zhou [15] ได้ทำการปรับปรุงวิธีการ Item-based ด้วยเทคนิคการจัดกลุ่ม ทำการเปรียบเทียบกับวิธีการคัดกรองแบบพึ่งพาผู้ใช้ร่วม ผลปรากฏว่าการจัดกลุ่ม มีประสิทธิภาพมากกว่า ทั้งนี้อาจเป็นเพราะว่าได้มีการกลั่นกรองข้อมูลก่อนด้วยการจัดกลุ่มก่อนเข้าสู่กระบวนการประมวลข้อมูลส่งผลให้ประสิทธิภาพดีมากขึ้น

Tiwari และ Shrivastava [36] ได้ทำการปรับปรุงวิธีการแนะนำข้อมูลด้วยการจำแนกข้อมูลค่าคะแนน การจัดกลุ่มผู้ใช้ การจัดกลุ่มประเภทภาพยนตร์ ก่อนนำข้อมูลเข้าสู่กระบวนการแนะนำข้อมูล และทำการกำหนดค่าน้ำหนักในกระบวนการสุดท้าย ผลปรากฏว่าระบบแนะนำข้อมูลมีประสิทธิภาพมากขึ้น สอดคล้องกับงานวิจัยของ J.Gupta,J.Gudge[37] ได้ศึกษากรอบการทำงานระบบแนะนำข้อมูล ด้วยวิธีการหาค่าความใกล้เคียงสินค้า และทำการจัดกลุ่มผู้ใช้ ก่อนเข้า

สู่กระบวนการแนะนำข้อมูล ทั้งนี้อาจเป็นเพราะว่าได้มีการกลั่นกรองข้อมูลก่อนด้วยการจัดกลุ่มก่อนเข้าสู่กระบวนการประมวลข้อมูล ส่งผลให้ประสิทธิภาพดีมากขึ้น

Codina และคณะ[38] ได้ศึกษาการกลั่นกรองข้อมูลเชิงความหมายและเพิ่มกระบวนการเตรียมข้อมูล (Pre-filtering) และกำหนดค่าเทรดโชว์ ก่อนเข้าสู่กระบวนการแนะนำข้อมูล ทำการเปรียบเทียบกับวิธีการ User-based Item-based การแยกตารางการคำนวณ การเตรียมข้อมูลแบบเดิม ผลปรากฏว่ามีประสิทธิภาพดีกว่า สอดคล้องกับงานวิจัยของ V.Codina และคณะ[39] ที่ทำการเพิ่มกระบวนการเตรียมข้อมูล ด้วยการเพิ่มขั้นตอน SVD และหาค่าความใกล้เคียงแบบโคไซน์

Ghazanfar และ Bennett [15] ได้ทำการปรับปรุงวิธีการระบบแนะนำข้อมูลแบบผสมผสานด้วยวิธีการสลับ โดยทำการหาค่าความถี่ กำหนดเทรดโชว์ หาความน่าจะเป็น และค่าความใกล้เคียง ทำการเปรียบเทียบกับวิธีการ User-based และ Item-based ผลปรากฏว่ามีประสิทธิภาพมากกว่า

Baltrunas และ Ricci [20] ได้ทำการแยกตารางการคำนวณจากข้อมูลค่าคะแนนด้วยวิธีการคัดกรองแบบพึ่งพาผู้ใช้ร่วม ทำการจำแนกข้อมูลด้วยวีธีการเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest neighbors (KNN) การแยกตารางคำนวณ หาค่าเฉลี่ย และทำการเปรียบเทียบกับวิธีการการแยกตารางข้อมูลสินค้าในการคำนวณ (Item-split) ผลปรากฏว่า Item-split มีประสิทธิภาพมากกว่า

Datta และคณะ[11] ได้จัดกลุ่มแยกตามเพศ-อายุ และเพศ-อาชีพ จากนั้นคำนวณค่าความใกล้เคียง ก่อนเข้าสู่กระบวนการแนะนำข้อมูล ผลปรากฏว่า การจัดกลุ่มตามเพศ-อายุมีประสิทธิภาพดีกว่าการจัดกลุ่มตามเพศ-อาชีพ

สรุปจากตาราง 3 ข้อมูลนำเข้าที่เป็นค่าคะแนน (rating) เป็นค่าคะแนนความรู้สึกของผู้ใช้ในการประเมินความชอบต่อรายการ ซึ่งส่วนมากนักวิจัยใช้ข้อมูลค่าคะแนน ส่วนข้อมูลของผู้ใช้ประกอบด้วยอายุ เพศ อาชีพ เมื่อนักวิจัยนำเข้าเข้ามาสามารถที่ช่วยแก้ไขปัญหาผู้ใช้ใหม่ได้ ข้อมูลคุณลักษณะรายการภาพยนตร์ ประกอบด้วยประเภท ผู้กำกับ นักแสดง สามารถช่วยแก้ไขปัญหาภาพยนตร์ใหม่ได้และมีประสิทธิภาพมากยิ่งขึ้น และการกลั่นกรองข้อมูลก่อนเข้าสู่กระบวนการประมวลผลข้อมูลส่งผลให้ประสิทธิภาพการทำงานมากขึ้น ส่วนข้อมูลบริบท ประกอบด้วยคีย์เวิร์ด แท็ก เวลา เพื่อน และเงื่อนไขอื่นๆ เมื่อนักวิจัยใช้ข้อมูลนี้ ผู้วิจัยเห็นว่าไม่สามารถช่วยเพิ่มประสิทธิภาพได้และประสิทธิภาพไม่แตกต่าง รวมทั้งต้องใช้เวลาในการประมวลผลมากขึ้นตามไปด้วย ดังนั้น ผู้วิจัยจึงเลือกเฉพาะข้อมูลนำเข้าที่มีความจำเป็นต่อนำไปแก้ไขปัญหาตามวัตถุประสงค์ของการ

วิจัย โดยจะนำข้อมูลที่เป็นค่าคะแนน อายุ เพศ อาชีพ ประเภทภาพยนตร์ ผู้กำกับ นักแสดง ประกอบด้วยนักแสดงชาย และนักแสดงหญิง

2.5.2 การเตรียมข้อมูล

การจัดเตรียมข้อมูล (Data Preparation) ข้อมูลที่อยู่ในฐานข้อมูล หรือระบบสารสนเทศ ส่วนใหญ่เป็นข้อมูลที่ไม่ได้ผ่านกระบวนการในการจัดการข้อมูล อาจมีข้อมูลที่ไม่สมบูรณ์และข้อมูลอาจมีความผิดปกติ ได้แก่ เขตข้อมูลบางตัวอาจมีความไม่ทันสมัย หรือซ้ำซ้อน ข้อมูลอาจมีการสูญหาย (Missing) ข้อมูลมีความผิดปกติ (Outlier) ข้อมูลไม่อยู่ในรูปแบบที่สามารถนำมาวิเคราะห์ในกระบวนการวิเคราะห์ข้อมูล หรือค่าข้อมูลอาจมีความไม่สม่ำเสมอ หรือไม่สอดคล้องกับกฎเกณฑ์ที่ตั้งเอาไว้ หรืออาจเป็นค่าที่ไม่สมเหตุสมผลตามความเป็นจริง ดังนั้นจึงได้มีการกลั่นกรองข้อมูลก่อนเข้าสู่กระบวนการต่อไป เพื่อให้ได้ข้อมูลที่ถูกต้องมากที่สุด และลดเวคเตอร์ให้มีขนาดเล็กลง ทำให้มีข้อมูลมีความชัดเจนและช่วยให้การประมวลผลในขั้นตอนต่อไปเร็วยิ่งขึ้น ซึ่งข้อมูล MovieLens เป็นข้อมูลที่อยู่ในฐานข้อมูล ข้อมูลที่ประกอบด้วย UserID MovieID และ rating ต้องแปลงข้อมูลในรูปแบบของเมทริกซ์ [42] [43] [34] ส่วนข้อมูล IMDB เป็นลักษณะข้อมูลที่เป็นข้อความไม่มีรูปแบบโครงสร้าง ดังนั้นงานวิจัย[15] [27] ต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบของเมทริกซ์ก่อน

การประมาณค่ากลุ่ม เป็นการช่วยให้การจัดกลุ่มได้จำนวนกลุ่มที่เหมาะสม ช่วยในการลดเวลาสุ่มจำนวนกลุ่ม (k) ซึ่งวิธีการประมาณค่าในการจัดกลุ่ม [44] [45]สามารถทำได้ดังนี้

1. คำนวนการจัดกลุ่ม เช่น ฟัซซี่ซีมีน หรือ เคมีน แล้วเพื่อให้เห็นความแตกต่างระหว่างค่า k ที่กำหนด เช่น กำหนดค่า k=10
2. สำหรับค่า k เป็นการคำนวนผลรวมทั้งหมดภายในกลุ่มยกกำลังสอง คือการวัดความแปรปรวนของข้อมูลภายในกลุ่ม
3. สร้างกราฟเส้น จากการคำนวนในข้อ 1.2
4. กราฟเส้นจะมีจุดหักเห ซึ่งเป็นจุดที่ชี้ให้เห็นความแตกต่างในการจัดกลุ่ม

2.5.3 การประมวลผลข้อมูล

2.5.3.1 การคัดกรองข้อมูลแบบอิงเนื้อหา

Philip Shola และ John [17] ได้พัฒนาระบบแนะนำข้อมูลเอกสารวิจัยจากห้องสมุดด้วยเทคนคิการคัดกรองข้อมูลแบบอิงเนื้อหา โดยอาศัยข้อมูลพฤติกรรมของผู้ใช้ที่ทำการดาวน์โหลด เปิด หรือกดชอบเอกสารวิจัยในระบบ แล้วทำการดูข้อมูลที่เป็นคีย์เวิร์ด ชื่อเรื่อง

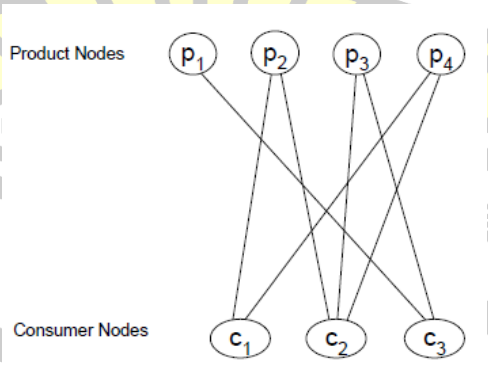
บทคัดย่อ สถานที่วิจัย เลขอ้างอิงงานวิจัย และผู้แต่ง มาทำการหาค่าความถี่ TF-IDF และค่าความใกล้เคียงแบบ Cosine ผลปรากฏว่าสามารถแนะนำข้อมูลงานวิจัยได้อย่างมีประสิทธิภาพ

Li และ Kim [26] ได้ศึกษาในการแก้ไขปัญหา Cold-Start ซึ่งได้นำข้อมูลคุณลักษณะของภาพยนตร์ประกอบด้วย นักแสดงชาย นักแสดงหญิง ผู้กำกับและประเภท และข้อมูลของผู้ใช้งานประกอบด้วย อายุ เพศ และอาชีพ และค่าคะแนนจากการดูภาพยนตร์ Lam และ คณะ (4) ได้นำข้อมูลประวัติของผู้ใช้งานประกอบด้วย อายุ เพศ อาชีพ และค่าคะแนนคะแนนจากการดูภาพยนตร์โดยใช้ฐานข้อมูลจาก MovieLens Li และ Kim ใช้วิธีการจัดกลุ่มจากคุณลักษณะของภาพยนตร์และผู้ใช้งาน จากนั้นทำการคำนวณค่าความใกล้เคียง ผลจากการทดลองแสดงให้เห็นว่าสามารถแก้ไขปัญหาได้

Gong [46] ได้ศึกษาการเรียนรู้โมเดลความสนใจของผู้ใช้สำหรับเนื้อหาที่ใช้ในการกรองส่วนบุคคลของระบบแนะนำข้อมูล เพื่อแก้ไขปัญหาที่มีข้อมูลขนาดใหญ่ ได้นำเสนอกรอบการทำงานของ Content-based ในการวิเคราะห์การสกัดเอกสารข้อมูล เวกเตอร์เอกสาร โมเดลความสนใจของผู้ใช้ วิธีการ Matching และการแก้ไขข้อมูลย้อนกลับของผู้ใช้ ทำให้สามารถแนะนำข้อมูลได้อย่างมีประสิทธิภาพ

2.5.3.2 การคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม

Huang Zeng และ Chen [47] ได้ศึกษาเรื่องการวิเคราะห์การเชื่อมโยงของระบบแนะนำข้อมูลภายใต้ข้อมูลเบาบาง เพื่อแก้ไขปัญหาข้อมูลเบาบาง แก้ไขปัญหาด้วยวิธีการการเชื่อมของกราฟระหว่างลูกค้ากับผลิตภัณฑ์ โดยได้ทดลองกับร้านหนังสือออนไลน์



รูปที่ 5 กราฟแสดงการเชื่อมโยงระหว่างลูกค้ากับผลิตภัณฑ์

เมื่อได้กราฟสามารถวิเคราะห์การเชื่อมโยงมาเป็นลักษณะของเมทริกซ์ได้ดังนี้

$$\begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \end{array} \begin{array}{c} \begin{array}{cccc} p_1 & p_2 & p_3 & p_4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \end{array}.$$

โดยค่าที่เป็น 0 คือค่าเชื่อมโยงระหว่างลูกค้ากับผลิตภัณฑ์ไม่ได้เชื่อมโยงกัน ส่วนค่า 1 คือการเชื่อมโยงระหว่างลูกค้ากับผลิตภัณฑ์ จากนั้นทำการคำนวณเมทริกซ์ด้วยวิธีการวิเคราะห์การเชื่อมโยง (Link Analysis) ผลการศึกษาพบว่าวิธีการ Link Analysis มีประสิทธิภาพดีกว่าวิธีการ User-based และวิธีการ Item-based

Ma [48] ได้นำเสนอเรื่อง ทำนายข้อมูลที่หายไปที่มีประสิทธิภาพสำหรับการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (CF) เพื่อแก้ไขข้อมูลเบาบาง ข้อมูลที่ใช้เป็น MovieLens จำนวน 100,000 เรคคอร์ด ผู้ใช้จำนวน 943 คน และภาพยนตร์จำนวน 1,682 เรื่อง ข้อมูลนำเข้าคือ คะแนนค่าคะแนน (1-5) ในการทดลองได้แบ่งจำนวนผู้ใช้เป็น 3 กลุ่ม เพื่อใช้สำหรับเป็นการเทรดข้อมูลคือ จำนวน 100 คน 200 คน และ 300 คน ทำการกำหนดค่า Lambda Eta และ Theta ที่มีค่า 0 และ 1 ที่แตกต่างกัน เพื่อทำการเปรียบเทียบอัลกอริธึม ดังนี้

user-based using PCC (UPCC)

the *item-based* methods (IPCC)

the effective missing data prediction (EMDP)

ผลปรากฏว่า EMDP มีประสิทธิภาพในการแก้ไขปัญหาเบาบางดีกว่าวิธีการ UPCC และ IPCC

Luo และคณะ [49] ได้นำเสนอเรื่องกรอบการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมบนพื้นฐานของความคล้ายคลึงกันของผู้ใช้ท้องถิ่น (Local) และความคล้ายคลึงกันของผู้ใช้ทั่วโลก (Global) ข้อมูลนำเข้าคือ คะแนนค่าคะแนน (1-5) จากฐานข้อมูล MovieLens โดยมีวัตถุประสงค์ดังนี้

1) เสนอวิธีการใหม่ (Surprisal-base Vector Similarity : SVS) เพื่อคำนวณความคล้ายคลึงกันผู้ใช้ท้องถิ่น (Local)

2) ใช้ระยะ Maximin ในการจับความสัมพันธ์ระดับโลก (Global) ของผู้ใช้ในการแก้ไขปัญหาข้อมูล sparsity

Maximin คือ ผลลัพธ์ที่ต่ำที่สุดในทุกทางเลือก แล้วเลือกค่าที่มากที่สุด

3) กรอบความร่วมมือการกรอง (LS & GS) จะเสนอให้อยู่บนพื้นฐานของความคล้ายคลึงกันผู้ใช้ท้องถิ่นและความคล้ายคลึงกันของผู้ใช้ทั่วโลก

วิธีการดำเนินการวิจัย ดังนี้

1) ได้ทำการ เปรียบเทียบอัลกอริธึม สำหรับผู้ใช้ Local ด้วยวิธีการ

PCC (Pearson Correlation Coefficient)

PCCS (Pearson Correlation Coefficient with significance weighting)

SVS (surprisal-based vector similarity) เป็นการคำนวณด้วยวิธี Vector Space Similarity จากความคะแนนของผู้ใช้ Local

SVSS (surprisal-based vector similarity with significance weighting)

2) เปรียบเทียบกรอบการทำงาน CF กับวิธีการอื่นๆ

user-based using PCC (UPCC)

the *item-based* methods (IPCC)

the similarity fusion algorithm (SF)

the effective missing data prediction (EMDP)

ผลปรากฏว่า วิธีการ SVSS มีประสิทธิภาพมากที่สุด โดยต้องมีคะแนนจากผู้ใช้ 20 คน วิธี EMDP มีประสิทธิภาพมากว่าวิธีการแบบดั้งเดิม (User-base) และ วิธี Global User Similarity ไม่สามารถปรับปรุงความถูกต้องได้ เมื่อมีคะแนนจากผู้ใช้ที่มีความหนาแน่น

Yildirim และ Krishnamoorthy [13] นำเสนอวิธีการเดินแบบสุ่มสำหรับการบรรเทาปัญหาเบาบาง ในการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม ข้อมูลในการทดลองเป็น MovieLens ซึ่งเป็นข้อมูลในการดูภาพยนตร์ จำนวน 1,00,2009 เรคคอร์ด ผู้ใช้จำนวน 6,040 คน ภาพยนตร์จำนวน 3,952 เรื่อง วัตถุประสงค์เพิ่มวิธีการคำนวณค่าใกล้เคียงกัน (Similarity) ในการแก้ไขปัญหาข้อมูลเบาบาง ข้อมูลนำเข้าเป็นรายละเอียดของข้อมูลภาพยนตร์ ซึ่งประกอบด้วย นักแสดง ผู้กำกับ โดยใช้การคำนวณค่าความคล้ายคลึงกันระหว่าง item ด้วย Cosine based Similarity และ Adjusted-Cosine Similarity ผลการศึกษาพบว่า การคำนวณความคล้ายคลึงกันด้วยวิธี Adjusted-Cosine Similarity ไปประยุกต์ใช้ เนื่องจากมีประสิทธิภาพดีกว่า Cosine based Similarity แต่ยังไม่สามารถตอบได้ว่าวิธีการคำนวณความคล้ายคลึงกันวิธีการไหนดีที่สุด โดยต้องทำการไปเปรียบเทียบกับวิธีการอื่นด้วย

Yang Zhang และ Wang [16] ได้นำเสนอเรื่องวิธีการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (CF) ปรับปรุงการปรับขยายของข้อมูลและการทำนายการเชื่อมโยงท้องถิ่น จากปัญหาการ

ขยายตัวของข้อมูลและข้อมูลบางเบา ข้อมูลที่ใช้เป็น MovieLens และ Netflix ซึ่งเป็นเว็บไซต์สำหรับการดูภาพยนตร์ออนไลน์ วิธีการดำเนินการวิจัยใช้ค่าความใกล้เคียงจากกราฟของรายการซึ่งคำนวนด้วยวิธีการ Adjusted Cosine Similarity และใช้ข้อมูลความสัมพันธ์ส่วนท้องถิ่นในการค้นหาพฤติกรรมของผู้ใช้ เมื่อคำนวนค่าความใกล้เคียง ให้นำค่ามาเพิ่มในชุดเมทริกส์ เพื่อนำเมทริกส์มาเป็นชุดเทรนข้อมูล ผลปรากฏว่า วิธีการ Scalable Item-base Collaborative Filtering (SICF) เมื่อเปรียบเทียบกับวิธีการ ItemKNN และ UserKNN จะเห็นได้ว่าวิธีการแบบ SICF มีประสิทธิภาพสูงกว่าทั้งสองวิธี

จากงานวิจัยนี้ จะเห็นได้ว่า หากมีวิธีการนำค่าความใกล้เคียงมาเพิ่มในชุดข้อมูลสามารถแก้ไขปัญหาข้อมูลเบาบางได้

2.5.3.3 การคัดกรองข้อมูลแบบผสมผสาน

Ghazanfar และ Bennett [12] ได้ศึกษาวิธีการลดการขยายตัวของข้อมูล และความถูกต้องของระบบแนะนำข้อมูลด้วยวิธีผสมผสาน เพื่อแก้ไขปัญหาการขยายตัวของข้อมูล ข้อมูลเบาบาง และปัญหาการเริ่มต้นของผู้ใช้หรือรายการ ข้อมูลที่ใช้เป็น MovieLens จำนวน 100,000 เรคคอร์ด ผู้ใช้จำนวน 943 คน และภาพยนตร์จำนวน 1,682 เรื่อง วิธีการดังนี้ 1) ข้อมูลค่าคะแนนในการคำนวณด้วยวิธี Item-based คือหาค่าความใกล้เคียงของรายการที่มีความสัมพันธ์กัน 2) ข้อมูลคุณลักษณะของรายการภาพยนตร์ ประกอบด้วย คีย์เวิร์ด แท็ก ผู้กำกับ นักแสดงชายหญิง เตรียมข้อมูลด้วยวิธีการ TF-IDF และค้นหาค่าความใกล้เคียง 3) ข้อมูลเพศของผู้ใช้ วิธีการประมวลผลใช้การคำนวณค่าความใกล้เคียงจากทั้ง 3 วิธีการ โดยใช้ Adjusted Cosine ในการคำนวณจากค่าคะแนน ใช้วิธีการคำนวณค่าความใกล้เคียงด้วย Vector Similarity ของคุณลักษณะของรายการภาพยนตร์และเพศของผู้ใช้ จากนั้นนำจากมากที่สุดมาทำการถ่วงน้ำหนักและพยากรณ์แนะนำข้อมูล ซึ่งผลปรากฏว่าวิธีการนี้ชื่อว่า $Boosted_{RDF}$ มีประสิทธิภาพมากกว่าวิธีการอื่น ประกอบด้วย

1) User-based with default voting (DV) คำนวณด้วยวิธีการ Pearson correlation

2) Item-based คำนวณด้วยวิธีการ Adjusted Cosine Similarity

3) IDemo4 คำนวณด้วยวิธีการแบบผสมผสาน

4) Naive Bayes โดยใช้คุณลักษณะของ Item ในการจำแนกประเภท

5) Naive hybrid คำนวณค่าเฉลี่ยจาก Content-based และ User-based CF

6) Pazzani เป็นการคำนวณแบบผสมผสาน CF และ Content-based จากการกลั่นกรองข้อมูลผู้ใช้

7) Personality diagnosis เป็นการกลั่นกรองข้อมูลจากข้อมูลผู้ใช้

Yoshii และคณะ [50] ได้ศึกษา ไฮบริดระหว่าง CF และเนื้อหา (Content-based) แนะนำเพลง การใช้โมเดลความน่าจะเป็นด้วยความชอบแฝงของผู้ใช้ โดยมีวัตถุประสงค์เพื่อเพิ่มความถูกต้องของระบบแนะนำข้อมูล แนะนำศิลปินใหม่ และแก้ไขปัญหา new item ด้วยวิธีการ CF ใช้ค่าความใกล้เคียงกันด้วยสูตรของ Pearson ส่วน Content-based พิจารณาเนื้อหาของเพลงว่ามีความชอบหรือไม่ หากชอบให้คะแนนเป็น 1 หากไม่ชอบให้คะแนนเป็น 0 แล้วทำการคำนวณค่าความใกล้เคียงกัน จากนั้นพิจารณาความหมายแฝงจากประเภทของเพลง พิจาณาความน่าจะเป็นจะในการแนะนำเพลงด้วยวิธีการ EM อัลกอริธึม ซึ่งเป็นวิธีการประมาณแบบ Gaussians ผลการทดลอง พบว่าวิธีการแบบไฮบริดมีประสิทธิภาพดีกว่า CF และ Content-based

Salter และ Antonopoulos [51] ได้ศึกษาในการแก้ไขข้อมูลที่มาใหม่ โดยใช้ข้อมูลค่าคะแนน และคุณลักษณะของข้อมูลภาพยนตร์ประกอบด้วย นักแสดง ผู้กำกับ และประเภท โดยใช้ฐานข้อมูลจาก MovieLens ผู้ใช้จำนวน 943 คน ภาพยนตร์ จำนวน 1682 เรื่อง ค่าคะแนนจำนวน 1 แสนค่าคะแนน วิธีการดังนี้

1) คำนวณค่าความใกล้เคียงกันค่าคะแนนที่มีความสัมพันธ์ของผู้ใช้แต่ละคน

2) จัดลำดับของภาพยนตร์

3) คำนวณค่าน้ำหนักเฉลี่ย

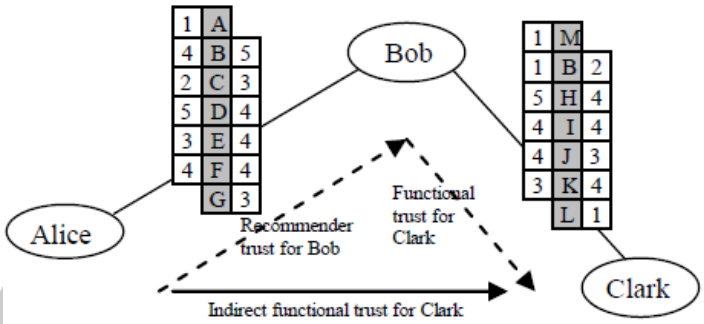
4) เพิ่มค่าน้ำหนักเฉลี่ยให้นักแสดง ผู้กำกับ และประเภท

5) คำนวณค่าเฉลี่ยคะแนนข้อมูลนักแสดง ผู้กำกับ และประเภท

6) คำนวณค่าเฉลี่ยของภาพยนตร์แต่ละเรื่อง

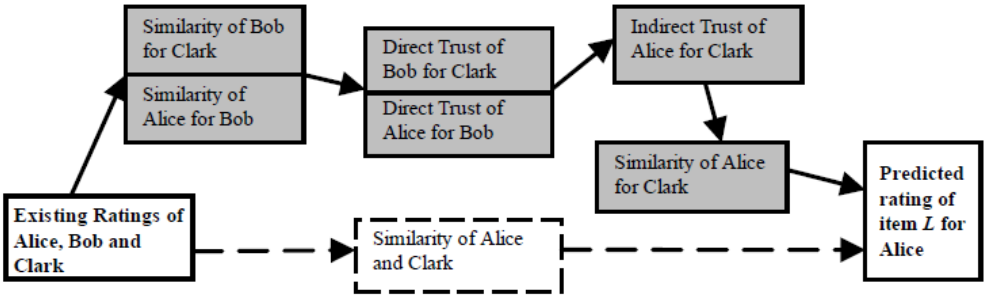
ผลจากกลั่นกรองแบบพึ่งพาผู้ใช้ร่วมแล้วกลั่นกรองเนื้อหามีประสิทธิภาพมากกว่ากลั่นกรองเนื้อหา กลั่นกรองแบบพึ่งพาผู้ใช้ร่วม กลั่นกรองเนื้อหาแล้วกลั่นกรองแบบพึ่งพาผู้ใช้ร่วม

Pitsilis และ Knapskog [35] ได้ศึกษาการเพิ่มความน่าเชื่อถือจากสังคมแก้ปัญหาข้อมูลเบาบางของระบบแนะนำข้อมูล ซึ่งข้อมูลที่ใช้เป็น MovieLens โดยมีวัตถุประสงค์เพิ่มประสิทธิภาพในระบบแนะนำข้อมูล และแก้ไขปัญหาข้อมูลเบาบางและการเริ่มต้นของผู้ใช้ วิธีการได้นำวิธีการลักษณะการวิเคราะห์ข้อมูลทางตรง (Direct) และข้อมูลทางอ้อม (Indirect) ข้อมูลนำเข้าคือค่าคะแนน (1-5)



รูปที่ 6 ลักษณะของข้อมูลทางตรงและทางอ้อม

เมื่อได้ข้อมูลทางตรงกับข้อมูลทางอ้อมจากความสัมพันธ์ของผู้ใช้ จะนำข้อมูลคะแนนมาคำนวณค่ามาใกล้เคียงกัน



รูปที่ 7 แนวคิดการคำนวณความใกล้เคียงกันของข้อมูลทางตรงและข้อมูลทางอ้อม

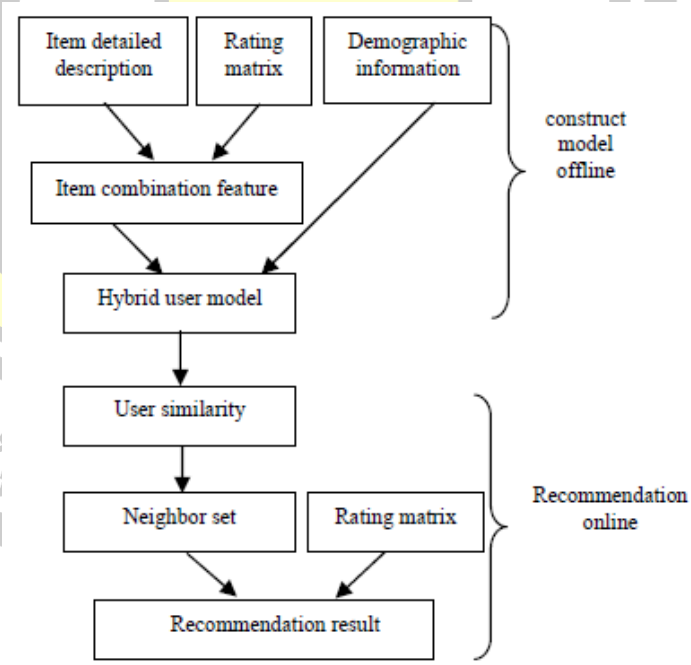
เมื่อคำนวณค่าความใกล้เคียงกันของข้อมูลทางตรงและข้อมูลทางอ้อมเพื่อเพิ่มความน่าเชื่อถือ แล้วทำการเปรียบเทียบกับวิธีการการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม พบว่าวิธีการแบบ Hybrid มีประสิทธิภาพดีกว่าแบบ CF

ข้อดี สามารถนำวิธีการการคิดคำนวณแบบทางตรงและทางอ้อมไปประยุกต์ในการคำนวณได้

Nazim Shretha และ Jo [52] ได้ศึกษาเรื่องปรับปรุงการกรองตามเนื้อหาโดยใช้การทำนายการคัดกรองแบบผู้ใช้ร่วมสำหรับแนะนำภาพยนตร์ เพื่อลดข้อจำกัดจากวิธีการกรองตามเนื้อหาและวิธีการคัดกรอกแบบผู้ใช้ร่วม จึงได้นำเสนอวิธีการแบบไฮบริด ข้อมูลที่ใช้ในการทดลองเป็น MovieLens วิธีการดำเนินการ การกรองข้อมูลเนื้อหาใช้ข้อมูลของนักแสดง ผู้กำกับ และประเภทของภาพยนตร์ ส่วนการกรองแบบผู้ใช้ร่วมใช้ข้อมูลของคะแนนจากผู้ใช้และคำนวณค่าความใกล้เคียงกันด้วยวิธีการของ Pearson Correlation ขั้นตอนวิธีการเลือกหลากหลายรายการ (DSA) ถูกนำมาใช้

เพื่อเลือกรายการ k หมู่รายการ แนะนำ โดยการกรองการทำงานร่วมกัน จะใช้ความแตกต่างกันที่รายการเป็นตัวชี้วัดเพื่อเลือกรายการเหล่านี้ ก็จะใช้เวลาหนึ่งรายการเป็นรายการที่ใช้งานกับสูงสุดที่คาดการณ์ไว้ คะแนนและคำนวณความแตกต่างกันกับส่วนที่เหลืออีก รายการที่มีคะแนนสูงสุดจะถูกนำใช้งาน ส่วนรายการที่เหลือจะถูกลบออกกระบวนการนี้ จะทำซ้ำแล้วซ้ำอีกจนกว่าเราจะได้ k รายการที่แตกต่างกัน การกรองใช้คะแนนผู้ใช้ที่ใช้งานอยู่และ k เลือกการประเมินรายการจากกระบวนการคัดเลือกหลากหลายรายการ (DSA) เป็นข้อูลนำเข้า จากนั้นจะทำการแนะนำข้อมูลต่อไป ซึ่งวิธีการนี้เรียกว่า Enhanced Content-based Filtering Using Diverse Collaborative Prediction (ECBDP) เมื่อนำมาเปรียบเทียบประสิทธิภาพกับวิธีการ Content-based Filtering (CBF) และ Naive Hybrid Approach (NHB) พบว่าวิธีการ ECBDP มีประสิทธิภาพดีกว่า

Wang Yuan และ Sun [53] ได้เสนอเรื่องการแนะนำข้อมูลด้วยวิธีการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (CF) บนพื้นฐานของรูปแบบผู้ใช้ผสมผสาน เพื่อแก้ไขปัญหาการขยายตัวของข้อมูล ข้อมูลคะแนนเบาบาง ข้อมูลที่ใช้เป็น MovieLens จำนวน 100,000 เรคคอร์ด ผู้ใช้จำนวน 943 คน และภาพยนตร์จำนวน 1,682 เรื่อง โดยมีวัตถุประสงค์เพื่อช่วยลดความซับซ้อนและระยะเวลารวม ปรับปรุงการขยายตัวของข้อมูล โดยสร้างรูปแบบผู้ใช้ใหม่และ CF ด้วยวิธีทางพันธุกรรมเรียนรู้รูปแบบน้ำหนักคุณลักษณะผู้ใช้ มีรูปแบบการทำงาน ดังนี้



รูปที่ 8 วิธีการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมบนพื้นฐานของรูปแบบผู้ใช้ผสมผสาน

วิธีการดำเนินงาน ทำการกลั่นกรองรายละเอียดของรายการข้อมูล ข้อมูลประชากรศาสตร์และคะแนนค่าคะแนน จากนั้นเข้าสู่กระบวนการ Hybrid User Model ซึ่งจะใช้หลักการ Feature Interest Measure (FIM) ประกอบด้วย

1) Total ค่าคะแนน (TR) ผลรวมของคะแนนของผู้ใช้กับรายการ

2) Feature ค่าคะแนน (FR) ผลรวมที่มีประสิทธิภาพของผู้ใช้กับรายการ

3) Feature Frequency (FF) ผลรวมระยะเวลาที่มีประสิทธิภาพ

4) Relative Feature ค่าคะแนน (RFR) ได้มาจาก FR/TR

5) Relative Feature Frequency (RFF) ได้มาจาก FF/TF

ผลปรากฏว่าวิธีการ Hybrid มีประสิทธิภาพสูงกว่าแบบ Content-based และ CF ข้อดี สามารถนำระยะเวลามาคำนวณ ในการเพิ่มความถูกต้องในการแนะนำ

Shinde [54] ได้ศึกษาเรื่องระบบแนะนำบุคคลแบบไฮบริดใช้วิธีการจัดกลุ่มแบบ K-medoids เพื่อแก้ไขการขยายตัวของข้อมูล โดยให้ข้อมูล Iris และ jester วิธีการจะดำเนินการ 2 กระบวน ลำดับแรกทำแบบออฟไลน์ ซึ่งจะค่าคะแนนของผู้ใช้และรายการมาทำการจัดกลุ่มด้วยเทคนิค K-medoids ลำดับที่สอง จะทำแบบออนไลน์ในการคำนวณค่าคะแนนมาเฉลี่ยน้ำหนักคำนวณค่าที่ใกล้เคียงด้วยวิธีการของ Pearson และแนะนำข้อมูลต่อ ผลการศึกษาพบว่า วิธีการ Fasst k-medoids มีประสิทธิภาพในเรื่องความเร็วกว่าวิธีการแบบ K-means K-medoids และ Fuzzy c-means

Braunhofer [28] ได้นำเสนอเทคนิคผสมผสาน สำหรับปัญหาเริ่มต้นรับรู้บริบทระบบแนะนำข้อมูล เพื่อแก้ไขปัญหาผู้ใช้ใหม่ รายการใหม่ และเนื้อหาใหม่ โดยใช้ข้อมูล STS CoMoDa และเพลง ซึ่งใช้เทคนิคแบบผสมผสาน ซึ่ง Context-Aware ผู้แนะนำระบบ (Carss) เป็นพิเศษประเภทของผู้แนะนำระบบ (RSS) ที่มีจุดมุ่งหมายที่ก่อให้เกิดคำแนะนำที่ถูกต้องมากขึ้น โดยไม่ได้ใช้ประโยชน์เฉพาะผู้ใช้แบบดั้งเดิมและขนาดรายการ แต่ยังเกี่ยวข้องข้อมูลบริบท (เช่น เวลา, สภาพอากาศ, สถานที่) ในขั้นตอนการเสนอแนะ วิธีการ Context-Aware Recommender systems (CARS) ได้รวมวิธีการแก้ไขปัญหา Cold-Staring ดังนี้

1) วิธีการ Context-Aware Matrix Factorisation for item categories (CAMF-CC) เป็นการนำค่าความแปรปรวนของคะแนนจากเมทริกซ์ของรายการ

2) วิธีการ SPF (Semantic Pre-Filtering) เป็นขั้นตอนวิธีการกรองล่วงหน้าด้วยปัจจัยจากเมทริกซ์การจัดอันดับแท็กทั้งหมดกับบริบทสถานการณ์ที่เหมือนกันหรือเทียบเท่า ซึ่งคาดว่าจะทำงานได้ดีในการแก้ไขปัญหา Cold-Start โดยใช้วิธีการ SVD

3) วิธีการ Content-based CAMF-CC เป็นนำค่าความแปรปรวนจากค่า CAMF-CC ที่มีความสัมพันธ์กับ item มาใช้เพื่อแก้ไขปัญหา New item โดยนำข้อมูลคุณลักษณะของ item เช่น ข้อมูลนักแสดง ประเภทของหนัง เป็นต้น

4) วิธีการ Demographics-based CAMF-CC เป็นนำค่าความแปรปรวนจากค่า CAMF-CC ที่มีความสัมพันธ์กับผู้ใช้ มาใช้เพื่อแก้ไขปัญหา New User โดยนำข้อมูลของผู้ใช้ เช่น ข้อมูลเพศ อายุ และลักษณะบุคลิกภาพ เป็นต้น

นอกจากนี้วิธีการ CARS ยังนำวิธีการน้ำหนักค่าเฉลี่ย (Average Weighted) วิธีการ Heuristic Switching และวิธีการปรับค่าน้ำหนัก (Adaptive Weighted)

ผลปรากฏว่า เมื่อนำวิธีการต่างๆ มาเปรียบเทียบเพื่อแก้ไขปัญหา new users พบว่าวิธีการปรับค่าน้ำหนัก (Adaptive Weighted) และวิธีการน้ำหนักค่าเฉลี่ย (Average Weighted) มีประสิทธิภาพดีที่สุด

แก้ไขปัญหา new items พบว่าวิธีการ Heuristic Switching มีประสิทธิภาพดีที่สุด

แก้ไขปัญหา new contexts พบว่าวิธีการปรับค่าน้ำหนัก (Adaptive Weighted) มีประสิทธิภาพดีที่สุด ตามมาด้วยวิธีการน้ำหนักค่าเฉลี่ย (Average Weighted) ซึ่งมีค่าน้ำหนักประสิทธิภาพไม่ต่างกันมากนัก

นอกจากนี้ Braunhofer Codina และ Ricci [27] ได้นำเสนอการวิธีการสลับแบบผสมผสานสำหรับปัญหาเริ่มต้นรับรู้บริบทระบบแนะนำข้อมูล โดยได้ทำวิธีการ Switching Hybrid (SHCA) เพื่อคำนวณคะแนนที่คาดการณ์โดยเฉลี่ยการคาดการณ์ของทั้งสองที่เป็นส่วนประกอบอัลกอริทึม ยกตัวอย่างเช่นการคาดการณ์สำหรับการจัดอันดับที่มีทั้งสำหรับผู้ใช้ใหม่และรายการใหม่เป็นค่าเฉลี่ยคะแนนประเมินโดยประชากรที่ใช้ CAMF-CC และเนื้อหาตาม CAMF-CC SHCA ดังที่ได้กล่าวไว้ใน (5) คาดว่าจะดีขึ้นรับมือกับทุกชนิดของสถานการณ์ Cold-Starting ที่พบ ระบบสามารถปรับบริบทให้เหมาะสมกับปัญหาที่เกิดขึ้น ผลปรากฏว่าวิธีการ SHCA เมื่อเปรียบเทียบประสิทธิภาพกับวิธีการ MF (Matrix Factorisation) วิธีการ CAMF-CC วิธีการ SPF Content-based วิธีการ CAMF-CC Demogr.-based CAMF-CC พบว่าวิธีการ SHCA มีประสิทธิภาพดีกว่าวิธีการอื่น

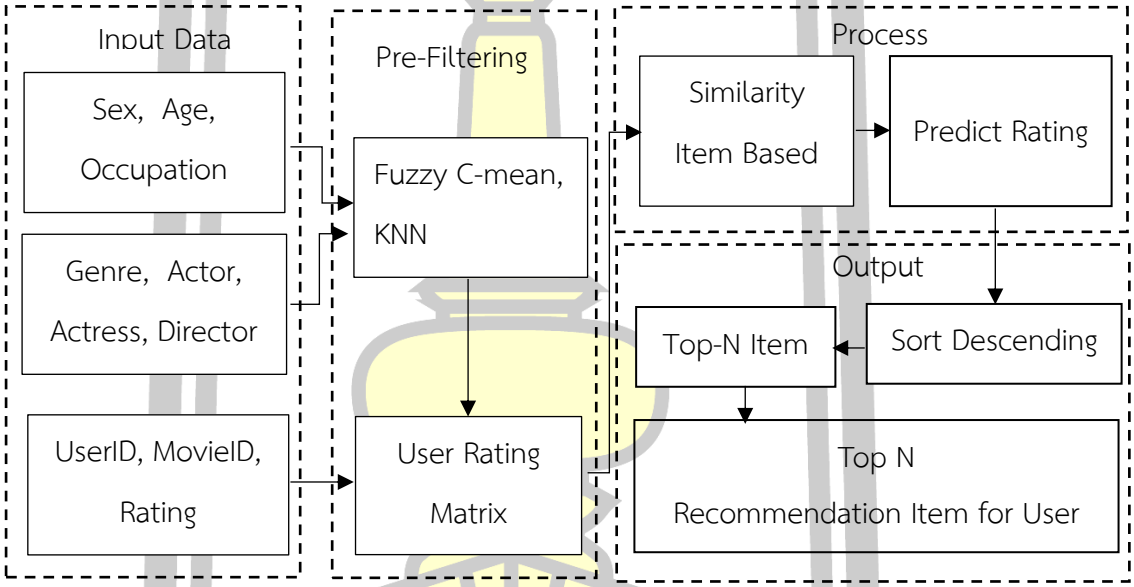
2.5.3.4 การวัดประสิทธิภาพ

การวัดประสิทธิภาพงานวิจัยของระบบแนะนำข้อมูล ซึ่งนักวิจัยหลายท่าน ได้มีการวัดประสิทธิภาพที่แตกต่างกัน อาทิเช่น Yuan-hong [55] มีการวัดประสิทธิภาพแบบ MAE และเวลา Liang และ Faqing [34] วัดประสิทธิภาพแบบ MAE และ Recall Gupta และ Gadge [37] วัดประสิทธิภาพแบบ MAE และ Coverage Nhat Lam และคนอื่น [10] วัดประสิทธิภาพแบบ

MAE และ NMAE Datta และคนอื่น [11] วัดประสิทธิภาพแบบ MAE RMSE Precision Recall F1 และเวลา Papagelis และ Plexousakis [33] วัดประสิทธิภาพแบบ MAE Pitsilis และ Knapskog [35] วัดประสิทธิภาพแบบ MAE และ F-Score Ghazanfar และ Prugel-Bennett [15] วัดประสิทธิภาพแบบ MAE Wu และคนอื่น [30] Bokde และคนอื่น[31] วัดประสิทธิภาพด้วยค่าความแม่นยำ สรุป จะเห็นได้ว่านักวิจัยโดยส่วนมากจะใช้การวัดประสิทธิแบบ MAE รองลงมาคือเวลา ดังนั้นผู้วิจัยเห็นว่าจะวัดประสิทธิภาพงานวิจัยแบบ MAE และความแม่นยำ

บทที่ 3
วิธีดำเนินการวิจัย

การดำเนินการวิจัยให้บรรลุวัตถุประสงค์ ผู้วิจัยได้อาศัยแนวคิด ทฤษฎี การวิจัยจากงานวิจัยที่เกี่ยวข้อง มีขั้นตอนวิธีการดำเนินการวิจัย 5 ขั้นตอน ได้แก่ 1) การรวบรวมข้อมูล 2) การวิเคราะห์ข้อมูล 3) การเตรียมข้อมูล 4) กระบวนการแนะนำข้อมูล 5) การวัดประสิทธิภาพ รายละเอียดดังต่อไปนี้



รูปที่ 9 ขั้นตอนการทำงานระบบแนะนำข้อมูล

จากรูปที่ 9 แสดงให้เห็นขั้นตอนการดำเนินการวิจัย ในงานวิจัยนี้ผู้วิจัยได้ทดลองในการแนะนำภาพยนตร์ นำข้อมูลมาจาก MovieLens และ IMDB โดยการวิจัยเริ่มจากกระบวนการเลือกข้อมูลที่เหมาะสม ทำการเตรียมข้อมูล แบ่งข้อมูลออกเป็น 3 กระบวนการ ประกอบด้วย 1) ข้อมูลคุณลักษณะของภาพยนตร์และบริบท 2) ข้อมูลเพศ อายุ ทำการแบ่งกลุ่มข้อมูล 3) ข้อมูลค่าคะแนนที่มีความสัมพันธ์ระหว่างผู้ดูภาพยนตร์กับเรื่องภาพยนตร์ด้วยการจำแนกประเภทข้อมูล จากนั้นเข้าสู่กระบวนการหาค่าความใกล้เคียง จัดเรียงข้อมูลค่าคะแนนจากมากไปน้อย และพยากรณ์แนะนำข้อมูลภาพยนตร์ที่มีค่ามากที่สุด

3.1 การรวบรวมข้อมูล

การรวบรวมข้อมูลจากที่ผู้วิจัยที่ได้ดำเนินการในบทที่ 2 หัวข้อที่ 2.2 และหัวข้อ 2.6.1 การวิเคราะห์ข้อมูล ซึ่งได้มีงานวิจัย [13] [32] [33] [34] ได้นำเสนอระบบแนะนำภาพยนตร์โดยข้อมูลนำเข้านักวิจัยส่วนใหญ่ใช้ข้อมูลค่าคะแนนในการนำข้อมูล นอกจากนี้มีข้อมูลผู้ใช้ประกอบด้วยอายุ เพศ อาชีพ ข้อมูลรายการภาพยนตร์ประกอบด้วยประเภท ผู้กำกับ นักแสดง ที่นำมาเป็นข้อมูลนำเข้าจากฐานข้อมูล MovieLens ไปใช้ในการทำวิจัย ชุดข้อมูลประกอบด้วยข้อมูลจำนวน 100,000 เร็คคอร์ด จากการเก็บรวบรวมข้อมูลผู้ใช้งาน จำนวน 943 คนและภาพยนตร์จำนวน 1,682 เรื่อง โดยชุดข้อมูลดังกล่าวสามารถดาวน์โหลดได้จากเว็บไซต์ http://grouplens.org/datasets/movielens/ ซึ่งข้อมูลมีรายละเอียดดังนี้

1. User Profile เป็นข้อมูลของผู้ใช้ ประกอบด้วย อายุ เพศ อาชีพ จำนวน 21 อาชีพ ประกอบด้วยอาชีพ administrator artist doctor educator engineer entertainment executive healthcare homemaker lawyer librarian marketing none other programmer retired salesman scientist student technician writer

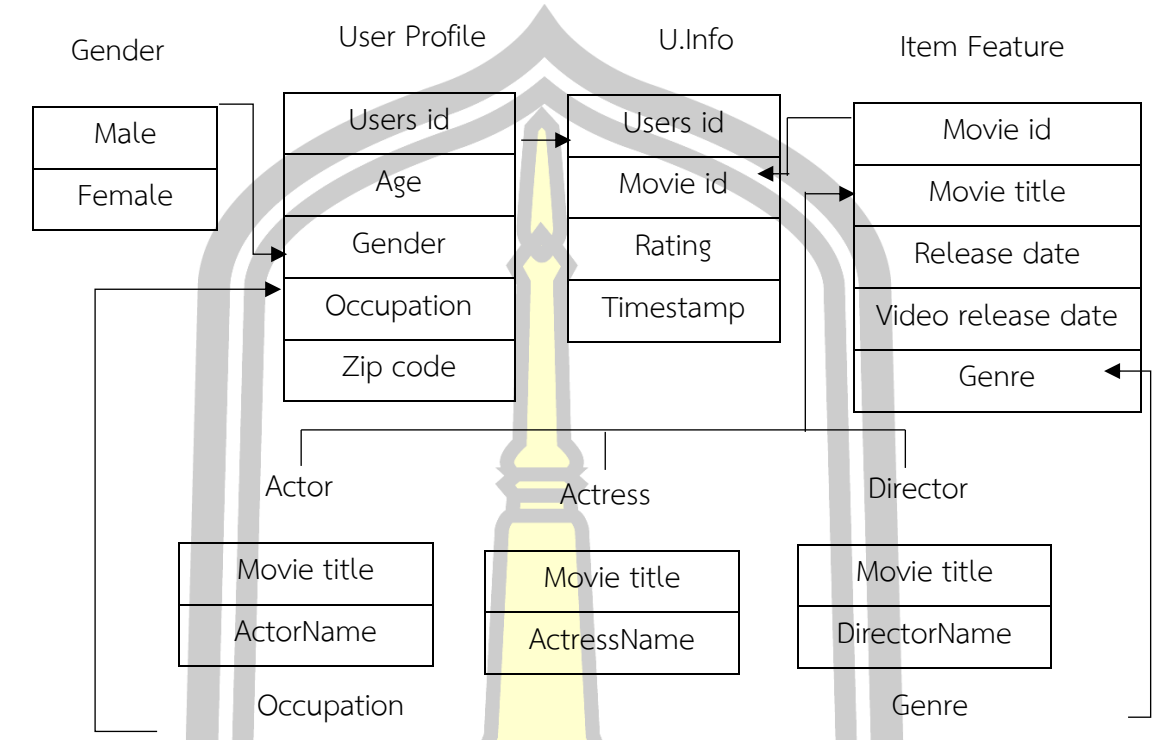
2. U.Info เป็นข้อมูลที่มีความสัมพันธ์ระหว่างผู้ใช้และภาพยนตร์ ประกอบด้วย รหัสผู้ใช้ รหัสภาพยนตร์ และค่าคะแนน (1-5)

3. Item Feature เป็นข้อมูลคุณลักษณะภาพยนตร์ ประกอบด้วย รหัสภาพยนตร์ ชื่อเรื่อง วันที่ฉาย วันที่เป็นวิดีโอ และประเภทภาพยนตร์ จำนวน 19 ประเภท คือ Action Adventure Animation Children's Comedy Crime Documentary Drama Fantasy Film-Noir Horror Musical Mystery Romance Sci-Fi Thriller War Western unknown

ต่อมาเพื่อเพิ่มประสิทธิภาพการทำงานระบบแนะนำข้อมูลนักวิจัย [12] [15] ได้นำข้อมูลจากฐานข้อมูล IMDB เฉพาะข้อมูลที่มีความสัมพันธ์กับข้อมูล MovieLens ทำการดาวน์โหลดข้อมูลเพิ่มเติม จากเว็บไซต์ http://www.imdb.com/ ซึ่งมีข้อมูลที่นำมาประกอบเพิ่มเติมคือ

1. ข้อมูลผู้กำกับ
2. ข้อมูลนักแสดงชาย
3. ข้อมูลนักแสดงหญิง

โดยลักษณะความสัมพันธ์ข้อมูล MovieLens และฐานข้อมูล IMDB ดังรูปที่ 10



Other or not specified	Lawyer
Academic/Educator	Programmer
Artist	Retired
Clerical/Admin	Sales/Marketing
Zip code	Scientist
Customer Service	Self-Employed
Doctor/Health Care	Technician/Engineer
Executive/Managerial	Tradesman/Craftsman
Farmer	Unemployed
Homemaker	Writer
Student	

unknown	Film-Noir
Action	Horror
Adventure	Musical
Animation	Mystery
Children's	Romance
Comedy	Sci-Fi
Crime	Thriller
Documentary	War
Drama	
Fantasy	

รูปที่ 10 แสดงความสัมพันธ์ของข้อมูล MovieLens และ IMDB

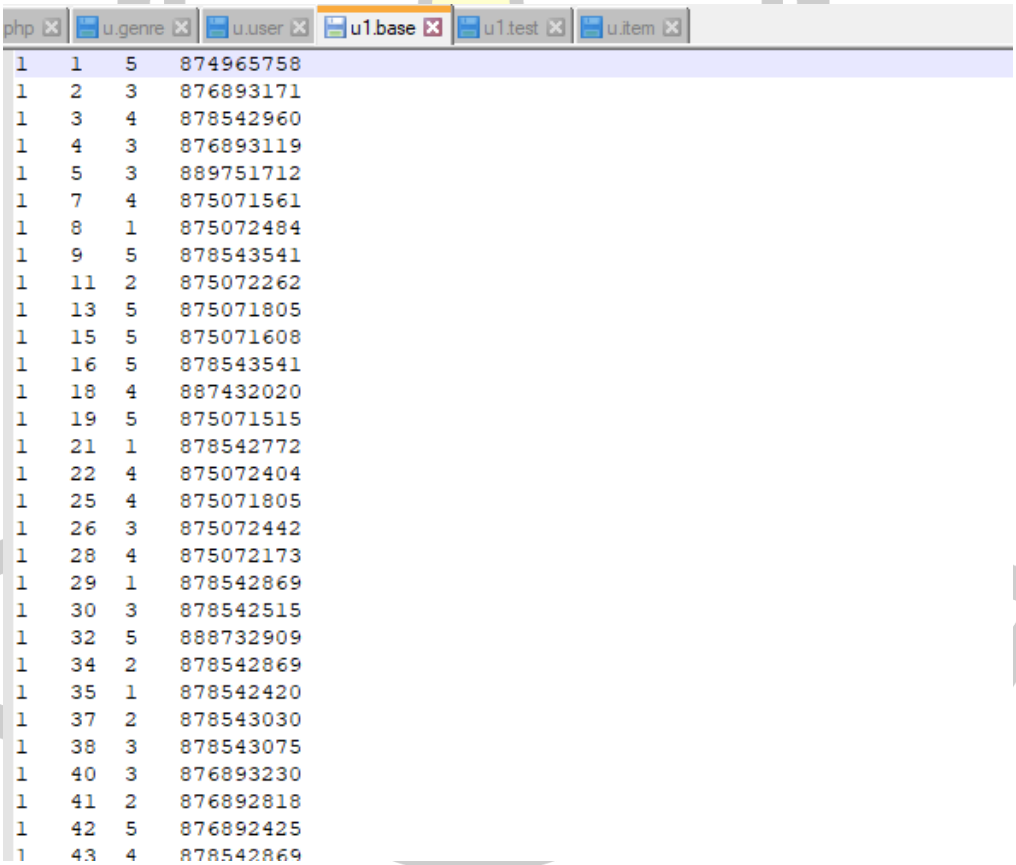
3.2 การจัดเตรียมข้อมูล

3.2.1 การแปลงข้อมูลข้อมูล

ข้อมูลนำเข้า ถือว่ามีความสำคัญอย่างยิ่งก่อนที่จะนำเข้าสู่กระบวนการจัดเตรียมข้อมูล ซึ่งผู้วิจัยได้ทำการวิเคราะห์ข้อมูลนำเข้าเพื่อให้ได้ข้อมูลที่มีความเหมาะสมในการแก้ไขปัญหาของระบบแนะนำข้อมูล

3.2.1.1 ข้อมูลค่าคะแนน

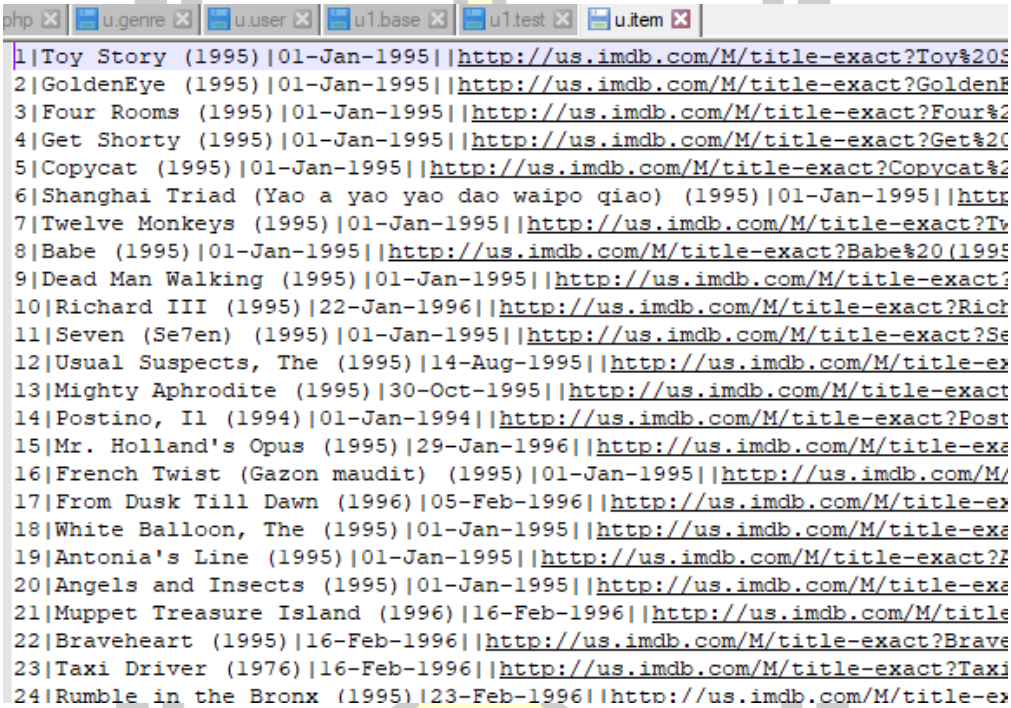
ข้อมูลค่าคะแนน (Rating) ในฐานข้อมูล MovieLens มีช่วงค่าคะแนนที่ 1-5 จำนวน 100,000 ค่าคะแนน จากข้อมูลผู้ใช้ 943 คน และภาพยนตร์ 1,682 เรื่อง [13] [32] [33] [34] ซึ่งเป็นไฟล์ข้อมูลที่ประกอบไปด้วย รหัสผู้ใช้ (userid) รหัสภาพยนตร์ (itemid) และค่าคะแนน (rating) และเวลา (Timestamp) ดังรูปที่ 11



รูปที่ 11 ข้อมูลค่าคะแนน

จากรูปที่ 11 มีงานวิจัย (6) (7) (3) ที่นำข้อมูลรหัสผู้ใช้ (userid) รหัสภาพยนตร์ (itemid) และค่าคะแนน (rating) ไปใช้ในการวิจัย ส่วนข้อมูลเวลา (Timestamp) ผู้วิจัยไม่นำไปใช้ เนื่องจากได้มีงานวิจัย He และWu [34] ได้นำมาใช้ในการหาค่าความถี่ ต้องใช้เวลาในการประมวลผลมากขึ้น และข้อมูลที่จัดเก็บเป็นข้อมูลที่มีค่าความถี่เป็นวันเดียวกันจำนวนมาก

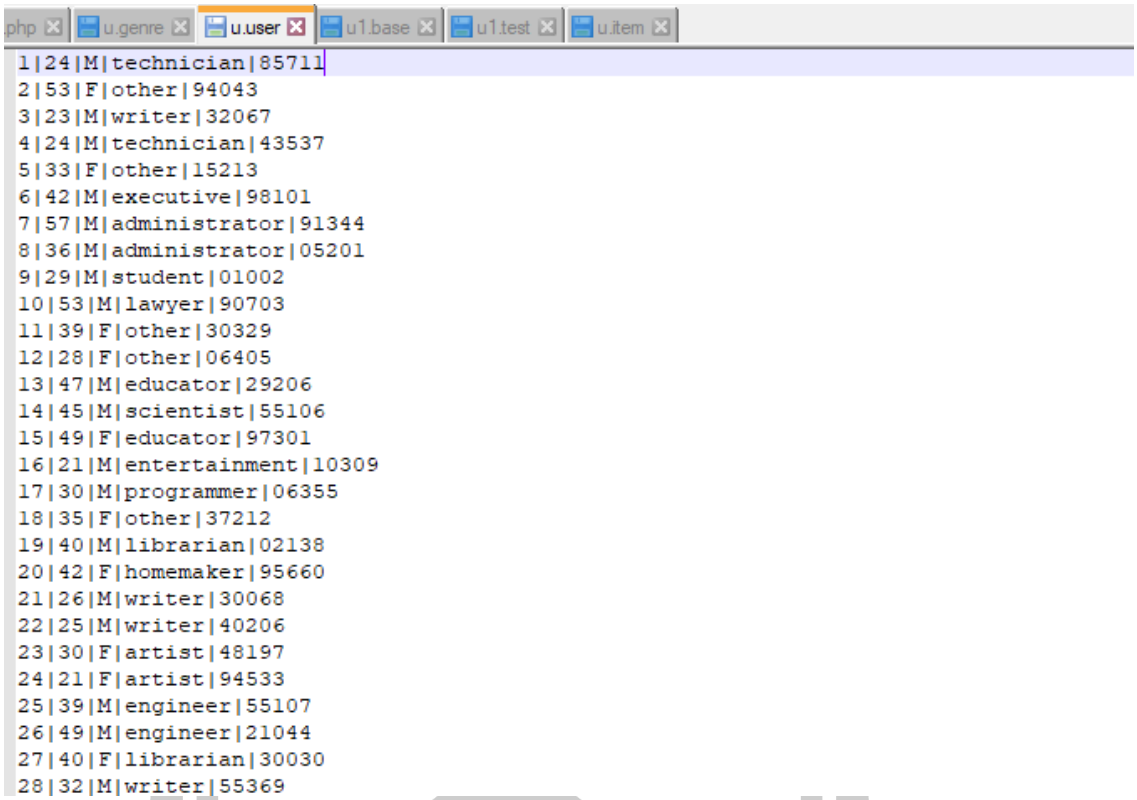
3.2.1.2 ข้อมูลคุณลักษณะประเภทภาพยนตร์ ประกอบด้วยข้อมูล Movieid Title Release date date Url และประเภทข้อมูลต่าง ๆ ดังรูปที่ 12



รูปที่ 12 ข้อมูลคุณลักษณะของภาพยนตร์

จากรูปที่ 12 ผู้วิจัยจะนำเฉพาะข้อมูล รหัสภาพยนตร์ (MovieId) ชื่อเรื่อง (MovieTitle) และประเภท (genre) ไปใช้ ส่วนแอตทริบิวต์ Release date และ URL ไม่นำไปใช้ในงานวิจัย

3.2.1.3 ข้อมูลคุณลักษณะของผู้ใช้ ประกอบด้วยข้อมูลรหัสผู้ใช้ (Userid) อายุ (Age) เพศ (Sex) อาชีพ (Occupation) และรหัสไปรษณีย์ (Code) ดังรูปที่ 13



```
1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43537
5|33|F|other|15213
6|42|M|executive|98101
7|57|M|administrator|91344
8|36|M|administrator|05201
9|29|M|student|01002
10|53|M|lawyer|90703
11|39|F|other|30329
12|28|F|other|06405
13|47|M|educator|29206
14|45|M|scientist|55106
15|49|F|educator|97301
16|21|M|entertainment|10309
17|30|M|programmer|06355
18|35|F|other|37212
19|40|M|librarian|02138
20|42|F|homemaker|95660
21|26|M|writer|30068
22|25|M|writer|40206
23|30|F|artist|48197
24|21|F|artist|94533
25|39|M|engineer|55107
26|49|M|engineer|21044
27|40|F|librarian|30030
28|32|M|writer|55369
```

รูปที่ 13 ข้อมูลคุณลักษณะผู้ใช้

จากรูปที่ 13 ผู้วิจัยจะทำการแปลงข้อมูลที่เป็นตัวอักษรให้เป็นตัวเลข ให้กับข้อมูลเพศ เช่น M=1 และ F = 2 และข้อมูลอาชีพ จะทำการนำข้อมูลอาชีพกำหนดเป็นแอตทริบิวต์ เช่น technician student ข้อมูลภายในเมทริกซ์จะกำหนดให้เป็นตัวเลข 1 หากข้อมูลของผู้ใช้เป็นอาชีพตามแอตทริบิวต์ และกำหนดให้เป็นเลข 0 หากข้อมูลของผู้ใช้ไม่เป็นอาชีพตามแอตทริบิวต์

3.2.1.4 ข้อมูลจากฐานข้อมูล IMDB ประกอบด้วย ข้อมูลผู้กำกับ ข้อมูลนักแสดงชาย และข้อมูลนักแสดงหญิง รายละเอียดดังนี้

1) ข้อมูลนักแสดงชาย ลักษณะข้อมูลจะเป็นเท็กซ์ไฟล์ (Text File) โดยมีการแยกชื่อเรื่องและชื่อนักแสดงชาย ผู้วิจัยได้ทำการแปลงข้อมูลให้อยู่ในตาราง ดังตาราง 4

ตาราง 4 ข้อมูลนักแสดงชาย

ลำดับ	ชื่อนักแสดงชาย	ชื่อเรื่องและปีที่ฉาย
1	Alexander, Keith (XXI)	Democracy Now! (2001)
2	Alexander, Keith (XXII)	Whistleblowers: The Untold Stories (2011)
3	Alexander, Kellen	Break-ups: Kellen & Seth (2010)
4	Alexander, Kelly (I)	Buried (2008/II)
5	Alexander, Ken (I)	God Squad! (2002)
6	Alexander, Ken (II)	Godchildren (1971)
7	Alexander, Kendell	Racing the Sunrise (2016),The Folklorist (2012)
8	Alexander, Kenneth (IV)	Comedy Fight Club (2007)
9	Alexander, Kenneth (V)	Frontline (1983)
10	Alexander, Kenneth Bam	Ellen: The Ellen DeGeneres Show (2003), Jimmy Kimmel Live! (2003)

จากตาราง 4 ข้อมูลนักแสดงชาย ผู้วิจัยทำการคัดกรองชื่อนักแสดงชายที่ไม่ซ้ำกันมา กำหนดเป็นแอตทริบิวต์ จากภาพยนตร์แต่ละเรื่องในฐานข้อมูล IMDB ดังตาราง 5

ตาราง 5 รายชื่อนักแสดงชายที่ไม่ซ้ำกัน

ลำดับ	ชื่อนักแสดงชาย
1	Alexander
2	Keith
3	Kellen
4	Kelly
5	Ken
6	Kendell
7	Kenneth
8	Kenneth Bam
9	Kenny
10	Kenrick

2) ข้อมูลนักแสดงหญิง ลักษณะข้อมูลจะเป็นเท็กซ์ไฟล์ (Text File) โดยมีการแยกชื่อเรื่องและชื่อนักแสดงหญิง ผู้วิจัยได้ทำการแปลงข้อมูลให้อยู่ในตาราง

ตาราง 6 ข้อมูลนักแสดงหญิง

ลำดับ	ชื่อนักแสดงหญิง	ชื่อเรื่องและปีที่ฉาย
1	Ancelet, Mary Caroline	Fripes de choix, guenilles de roi (1998)
2	Ancelin, Christine	Louise-Michel (2008)
3	Ancelin, Jennifer	Leapling
4	Ancelin, Luna	Duel au soleil (2014)
5	Ancell, Stephanie	Treasure (2012)
6	Ancelli, Marisa	Genitori in blue-jeans (1960)
7	Ancelot, Claudine	L"ombra abitata (1995) (TV)
8	Anceny, Dominique	La lisi\E8re (2010)
9	Ancer, Aurora	Burger P.I. (2004)
10	Ancer, Rosario	A Window Looking In (2010)
11	Anceschi, Rosalba	I migliori sentimenti (2007)
12	Anchal (II)	Uthsaham (2003)

จากตาราง 6 ผู้วิจัยทำการคัดกรองชื่อนักแสดงหญิงที่ไม่ซ้ำกันมากำหนดเป็นแอตทริบิวต์จากภาพยนตร์แต่ละเรื่องในฐานข้อมูล IMDB ดังตาราง 7

ตาราง 7 รายชื่อนักแสดงหญิง ที่ไม่ซ้ำกัน

ลำดับ	ชื่อนักแสดงหญิง
1	Ancelet
2	Mary Caroline
3	Christine
4	Jennifer
5	Ancelin
6	Luna

ตาราง 7 (ต่อ)

ลำดับ	ชื่อนักแสดงหญิง
7	Ancell
8	Anceny
9	Ancelli
10	Marisa

3) ข้อมูลผู้กำกับ ลักษณะข้อมูลจะเป็นเท็กซ์ไฟล์ (Text File) โดยมีการแยกชื่อเรื่องและชื่อผู้กำกับ ผู้วิจัยได้ทำการแปลงข้อมูลให้อยู่ในตาราง

ตาราง 8 ข้อมูลผู้กำกับ

ลำดับ	ชื่อผู้กำกับ	ชื่อเรื่องและปีที่ฉาย
1	Aamodt, Kitty	Alcohol By Volume (2012)
2	Aamodt, V. Blackhawk	Il mutilato (2009)
3	Aamuri, Dhanvignesh (I)	Ali"s Letter (2014)
4	Aanenson, Quentin	A Fighter Pilot"s Story (1993) (TV)
5	Aappli, Simon	24 Hours (2000) (TV)
6	Aardal, Brynjar Fausk	Hage Wars (1998)
7	Aarden, Simon	Zoals de bom thuis tikt... (1999) (TV)
8	Aarhus, Aslak	Ole Bull - Himmelstormeren (2006)
9	Aarma, Kiur	Kullaketrajad (2013)
10	Aarniala, Timo	The Who Suomessa (1967)

จากตาราง 8 ผู้วิจัยทำการคัดกรองชื่อผู้กำกับที่ไม่ซ้ำกันมากำหนดเป็นแอตทริบิวต์ จากภาพยนตร์แต่ละเครื่องในฐานข้อมูล IMDB ดังตาราง 9

ตาราง 9 รายชื่อผู้กำกับที่ไม่ซ้ำกัน

ลำดับ	ชื่อผู้กำกับ
1	Aamodt
2	Kitty

ตาราง 9 (ต่อ)

ลำดับ	ชื่อผู้กำกับ
3	V. Blackhawk
4	Aamuri
5	Dhanvignesh
6	Aanenson
7	Quentin
8	Aappli
9	Simon
10	Aarhus

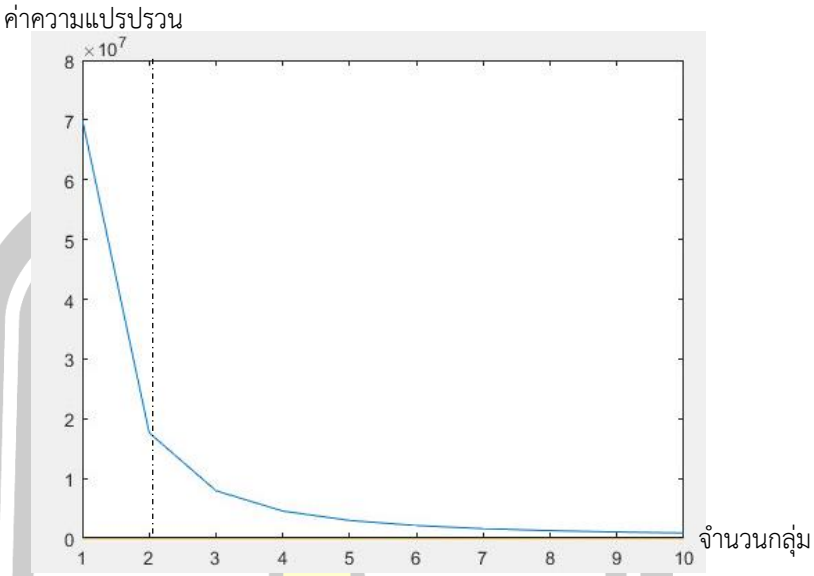
3.2.1.5 การประมาณค่ากลุ่ม (k) ซึ่งก่อนจะทำการจัดกลุ่ม ต้องการค่า k ของกลุ่มก่อน ซึ่งผู้วิจัยได้ทำการค้นหาค่า k ด้วยการหาค่าระยะห่างของข้อมูลด้วยกำหนดเป็นกราฟเส้น เพื่อดูความหักเหของเส้นกราฟ เมื่อได้ค่า k จึงทำการจัดกลุ่มของผู้ใช้ หรือข้อมูลลักษณะของภาพยนตร์ สามารถทำได้ดังนี้

1) คำนวนการจัดกลุ่ม เช่น ฟัซซี่ซีมีน หรือ เคมีน แล้วเพื่อให้เห็นความแตกต่างระหว่างค่า k ที่กำหนด เช่น กำหนดค่า k=10

2) สำหรับค่า k เป็นการคำนวนผลรวมทั้งหมดภายในกลุ่มยกกำลังสอง คือการวัดความแปรปรวนของข้อมูลภายในกลุ่ม

3) สร้างกราฟเส้น จากการคำนวนในข้อ 2)

4) กราฟเส้นจะมีจุดหักเห ซึ่งเป็นจุดที่ชี้ให้เห็นความแตกต่างในการจัดกลุ่ม ดังรูปที่ 14



รูปที่ 14 การหักเหของเส้นกราฟ

3.2.1.6 การจัดกลุ่มด้วยฟัซซี่ซีมีน หลักจากได้ค่าจากการประมาณค่ากลุ่ม ผู้วิจัยทำการแบ่งกลุ่มข้อมูลผู้ใช้ และข้อมูลลักษณะภาพยนตร์ ด้วยเทคนิค Fuzzy c-mean [25] [56] มีขั้นตอนการทำงาน ดังนี้

1) กำหนดกลุ่มข้อมูลที่ต้องการจัดกลุ่ม เพื่อกำหนดค่าเพื่อเป็นเงื่อนไขในการให้ข้อมูล หยุดการจัดกลุ่ม กำหนดค่าฟัซซี่พารามิเตอร์ (m) ซึ่งต้องมากกว่าหนึ่ง และ กำหนดจุดศูนย์กลางเริ่มต้นของข้อมูล

2) คำนวณค่าการเป็นสมาชิกของข้อมูลต่อกลุ่มข้อมูลต่างๆ

3) คำนวณจุดศูนย์กลางกลุ่มข้อมูลใหม่และตรวจสอบเงื่อนไขโดยตรวจสอบค่าการเป็นสมาชิกใหม่ลบค่าการเป็นสมาชิกก่อนหน้า

4) ถ้าเงื่อนไขเป็นจริงคำนวณค่าการเป็นสมาชิกและ Objective Function ถ้าเงื่อนไขเป็นเท็จ คำนวณค่าการเป็นสมาชิกจากจุดศูนย์กลางล่าสุด (วนรอบ)

3.2.1.7 การแบ่งกลุ่มข้อมูลคะแนน (1-5) และคำนวณค่าความใกล้เคียงวิธี Correlation Based Similarity แบบ Item-based [42] จากข้อมูลที่ประกอบด้วย รหัสผู้ใช้ (Userid) รหัสภาพยนตร์ (Movieid) และค่าคะแนน (Rating) ซึ่งจะอาศัยความสัมพันธ์จากแอตทริบิวต์ของ Userid หรือ Movieid จากการจัดกลุ่มด้วยฟัซซี่ซีมีที่มีการแบ่งกลุ่มข้อมูลผู้ใช้ และข้อมูลลักษณะภาพยนตร์ ดังตาราง 10

ตาราง 10 ค่าคะแนนการดูภาพยนตร์

User/Movie	M1	M2	M3	M4	M5
u1	4	4	3	4	2
u2	0	2	0	4	1
u3	5	1	2	0	3
u4	5	0	4	0	5

จากตาราง 10 คำค่าคะแนนจากผู้ใช้มาหาค่าความใกล้เคียงด้วยวิธี Correlation Based Similarity แบบ Item-based

$$sim(u,v) = \frac{\sum_{i \in I}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I}(r_{v,i} - \bar{r}_v)^2}} \quad (18)$$

เมื่อ $sim(u,v)$ แทนค่าความคล้ายคลึงระหว่างชิ้นข้อมูล u กับ v

$r_{u,i}$ และ $r_{v,i}$ แทนคะแนนชิ้นข้อมูล u มีต่อชิ้นข้อมูล i และคะแนนชิ้นข้อมูล v ต่อชิ้นข้อมูล i

$\bar{r}_u$ และ $\bar{r}_v$ แทนค่าเฉลี่ยของชิ้นข้อมูล u และ v

$\sum i \in I$ แทนผลรวมของชิ้นข้อมูลทั้งหมด

ตัวอย่าง หาความสัมพันธ์ระหว่างภาพยนตร์ M1 กับ M2 จากตาราง 10 โดยก่อนอื่นให้หาค่าเฉลี่ยของค่าคะแนนทั้งหมดจาก M1 และ M2

$$\bar{r}_{M1} = \frac{4+0+5+5}{4} = 3.5$$

$$\bar{r}_{M2} = \frac{4+2+1+0}{4} = 1.75$$

เมื่อได้ค่าเฉลี่ย M1 เท่ากับ 3.5 และ M2 เท่ากับ 1.75 ให้นำข้อมูลค่าคะแนนมาแทนค่าตามสมการ (20) ได้ข้อมูลการแทนค่า ดังนี้

$$sim(M1,M2) = \frac{\{(4-3.5)x(4-1.75)\}+\{(0-3.5)x(2-1.75)\}+\{(5-3.5)x(1-1.75)\}+\{(5-3.5)x(0-1.75)\}}{\sqrt{(4-3.5)^2+(0-3.5)^2+(5-3.5)^2+(5-3.5)^2} \quad x \quad \sqrt{(4-1.75)^2+(2-1.75)^2+(1-1.75)^2+(0-1.75)^2}}$$

$$sim(M1, M2) = \frac{1.125 + (-0.875) + (-1.125) + (-2.625)}{4.123106 + 2.95804}$$

$$sim(M1, M2) = \frac{-3.5}{12.19631} = -0.28697$$

เมื่อทำการหาค่าความใกล้เคียงทุกคู่ของภาพยนตร์ จะได้ค่าคะแนนความใกล้เคียง ดังตาราง 11

ตาราง 11 ค่าความใกล้เคียงด้วยวิธีการ Item Based

Movies	M1	M2	M3	M4	M5
M1	0	-0.28697	0.860916	-0.72761	0.778924
M2	-0.28697	0	-0.2	0.845154	-0.71429
M3	0.860916	-0.2	0	-0.50709	0.828571
M4	-0.72761	0.845154	-0.50709	0	-0.84515
M5	0.778924	-0.71429	0.828571	-0.84515	0

จากตาราง 11 ค่า 0 หมายความว่า เป็นข้อมูลภาพยนตร์เรื่องเดียวกัน ค่าลบ หมายความว่าค่าความใกล้เคียงกันที่มีความสัมพันธ์ของระหว่างภาพยนตร์ทั้งสองเรื่องมีความสัมพันธ์กันเชิงลบ และ ค่าบวก หมายความว่าค่าความใกล้เคียงกันที่มีความสัมพันธ์ของระหว่างภาพยนตร์ทั้งสองเรื่องมีความสัมพันธ์กันเชิงบวก

3.2.2 การทำความสะอาดข้อมูล ผู้วิจัยได้ทำการลดข้อมูลที่ไม่ต้องการ คือลดจำนวนของค่าคะแนนที่เป็น 0 โดยใช้เกณฑ์ในการลดคือ หาค่าเฉลี่ยของภาพยนตร์ต่อผู้ใช้ ซึ่งมีค่าเท่ากับ 65 หมายถึงภาพยนตร์หนึ่งเรื่องต้องมีคนดูอย่างน้อย 65 คน ถ้าคนดูน้อยกว่านี้จะไม่นำมาเป็นข้อมูลในการทดลอง

3.3 ระบบแนะนำข้อมูล

ระบบแนะนำข้อมูล สำหรับเพื่อแก้ไขปัญหาผู้ใช้ใหม่ และแก้ไขปัญหาภาพยนตร์ใหม่ ผู้วิจัยได้นำข้อมูลและวิธีการในการแก้ไขปัญหา รายละเอียดดังนี้

3.3.1 การแก้ไขปัญหาผู้ใช้ใหม่ มีขั้นตอนการทำงาน ดังนี้

3.1.1.1 ข้อมูลนำเข้า ได้แก่ ข้อมูลของผู้ใช้ (User Profile) ที่ประกอบด้วยข้อมูลเพศ (Sex) อายุ (Age) อาชีพ (Occupation) และข้อมูลการให้ค่าคะแนนจากผู้ใช้ในการดูภาพยนตร์ ประกอบด้วย รหัสผู้ใช้ (Userid) รหัสภาพยนตร์ (Movieid) และค่าคะแนน (Rating) จากฐานข้อมูล MovieLens

3.1.1.2 การประมวลผลด้วย Content Based ในกรณีที่มีผู้ใช้ใหม่ซึ่งไม่มีข้อมูลค่าคะแนนในการดูภาพยนตร์ จะทำการนำข้อมูลของผู้ใช้ใหม่ประกอบด้วย เพศ อายุ อาชีพ วัดระยะห่างจากข้อมูลจุดศูนย์กลางของกลุ่ม (Centroid) เพื่อทำให้ทราบข้อมูลผู้ใช้จะเข้าไปเป็นสมาชิกในกลุ่มใดกลุ่มหนึ่ง เมื่อเข้าไปยังกลุ่มที่เตรียมไว้ จากนั้นจะทำการหาค่าความใกล้เคียงของผู้ใช้ภายในกลุ่ม ด้วยเทคนิค KNN ซึ่งจะกำหนดค่าเทรดโชว์ และกำหนดค่า K=12 จะไปทำตามขั้นตอนถัดไป

3.1.1.3 การประมวลผลด้วย Collaborative Based เพื่อแนะนำข้อมูลภาพยนตร์ให้กับผู้ใช้ เป็นนำข้อมูลค่าคะแนน และข้อมูลความใกล้เคียงที่เตรียมไว้แล้ว ดังตาราง 10 และตาราง 11 นำมาหาค่าน้ำหนักรวมของข้อมูล โดยผู้วิจัยทำการหาค่าน้ำหนักรวมจำนวนตามที่ได้จากหาค่า KNN โดยทำทีละคนจนครบตามจำนวนคำนวณไว้ ซึ่งจะนำข้อมูลที่มีค่าน้ำหนักมากจำนวน 10 เรื่องของแต่ละคนมารวมกันแล้วจัดเรียงจากมากไปน้อย แล้วนำข้อมูลภาพยนตร์ตามจำนวนที่เหมาะสมไปแนะนำ ซึ่งค่าน้ำหนักรวมของข้อมูล สามารถแทนค่าข้อมูลดังสมการ (19)

$$pred(u,i) = \frac{\sum_{j \in ratedItems(u)} itemSim(i,j).r_{uj}}{\sum_{j \in ratedItems(u)} |itemSim(i,j)|} \quad (19)$$

เมื่อ $itemSim(i,j)$ คือ ค่าความใกล้เคียงของชิ้นข้อมูล i กับ j

r_{uj} คือ ค่าคะแนนของผู้ใช้ u กับชิ้นข้อมูล j

<u>ตัวอย่าง</u> เมื่อต้องการพยากรณ์แนะนำข้อมูลภาพยนตร์ให้กับผู้ใช้ u1 ในการคำนวณหาค่าน้ำหนักจากภาพยนตร์แต่ละเรื่อง เพื่อหาค่าน้ำหนักของภาพยนตร์ที่มีน้ำหนักมากที่สุด ในการแนะนำข้อมูลภาพยนตร์ให้กับผู้ใช้ รายละเอียดดังนี้

$$pred(u1,m1) = \frac{(4x(-0.28697)) + (3x0.860916) + (4x(-0.72761)) + (2x0.778924)}{|(-0.28697)| + |0.860916| + |(-0.72761) + |0.778924|}$$

$$pred(u1,m1) = \frac{0.082276}{2.65442} \approx 0.030996$$

$$pred(u1,m2)=\frac{(4x(-0.28697))+(3x(-0.20))+(4x0.845154)+(2x(-0.71429))}{|(-0.28697)|+|(-0.20)|+|0.845154|+|(-0.71429)|}$$

$$pred(u1,m2)=\frac{0.204156}{2.046414}\approx 0.099763$$

$$pred(u1,m3)=\frac{(4x0.860916)+(4x(-0.20))+(4x(-0.50709))+(2x0.828571)}{|0.860916|+|(-0.20)|+|(-0.50709)|+|0.828571|}$$

$$pred(u1,m3)=\frac{2.272446}{2.396577}\approx 0.948205$$

$$pred(u1,m4)=\frac{(4x(-0.72761)+(4x0.845154)+(3x(-0.50709))+(2x(-0.84515))}{|(-0.72761)|+|0.845154)|+|(-0.50709)|+|(-0.84515)|}$$

$$pred(u1,m4)=\frac{-2.74139}{2.925004}\approx -0.93723$$

$$pred(u1,m5)=\frac{(4x(-0.778924)+(4x(-0.71429))+(3x0.828571)+(4x(-0.84515))}{|(-0.778924)|+|-0.71429)|+|0.828571|+|(-0.84515)|}$$

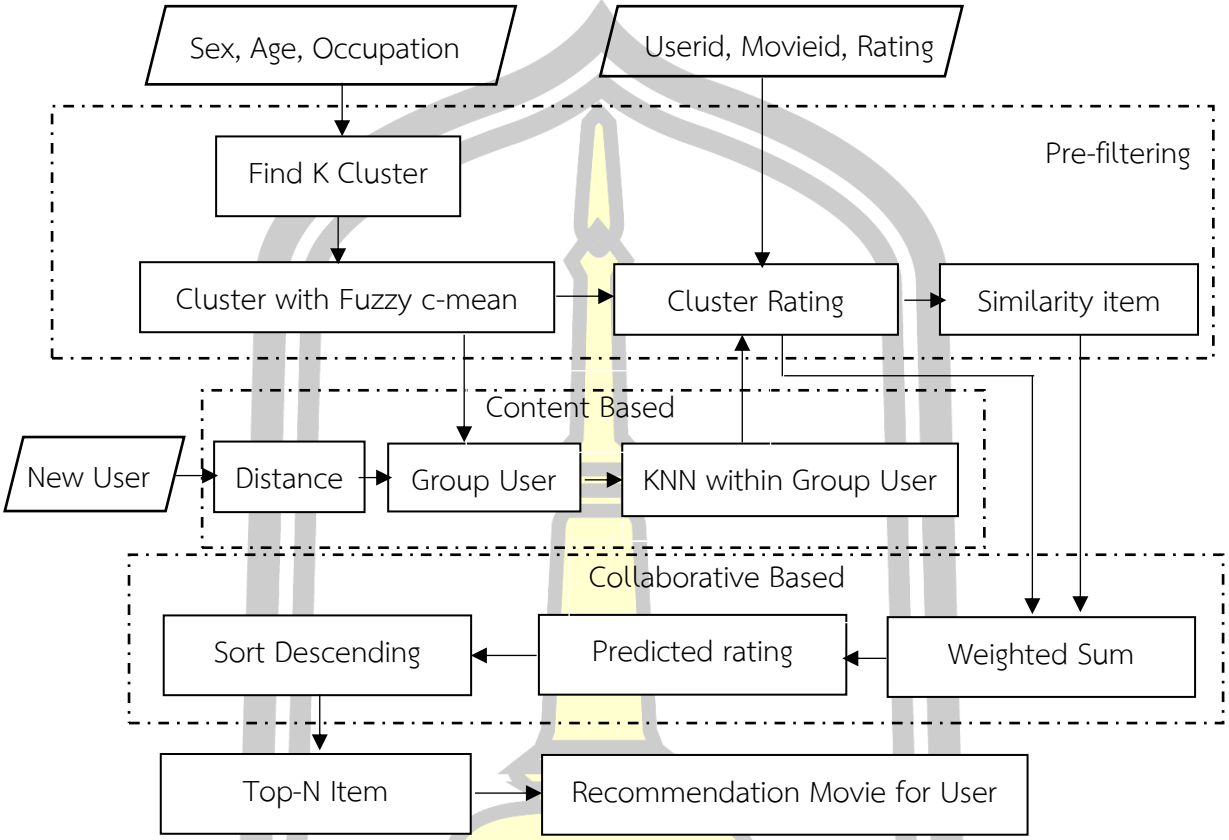
$$pred(u1,m5)=\frac{-6.86774}{3.166935}\approx -2.16858$$

โดยสรุปการหาค่าน้ำหนักรวมในการแนะนำข้อมูลภาพยนตร์ของผู้ใช้ u1 โดยเรียงลำดับจากมากไปน้อย ซึ่งการแนะนำข้อมูลจะเลือกค่าคะแนนที่มีน้ำหนักมากที่สุดไปแนะนำข้อมูลให้กับผู้ใช้ ดังตาราง 12

ตาราง 12 การเรียงลำดับข้อมูลภาพยนตร์ในการแนะนำข้อมูล

ลำดับที่	ชื่อภาพยนตร์	น้ำหนักค่าคะแนน
1	m3	0.948205
2	m2	0.099763
3	m1	0.030996
4	m4	-0.93723
5	m5	-2.16858

3.1.1.4 การแนะนำข้อมูลภาพยนตร์ให้กับผู้ใช้ใหม่ โดยจะนำข้อมูลจากตาราง 12 มาเลือกค่าคะแนนที่มีค่ามากที่สุด [39] [57] เพื่อแนะนำภาพยนตร์ให้กับผู้ใช้ต่อไป ทั้งนี้สามารถสรุปเป็นภาพรวมการทำงาน ดังรูปที่ 15



รูปที่ 15 ขั้นตอนการแก้ไขปัญหาผู้ใช้ใหม่

3.3.2 การแก้ไขปัญหาภาพยนตร์ใหม่ มีขั้นตอนการทำงาน ดังนี้

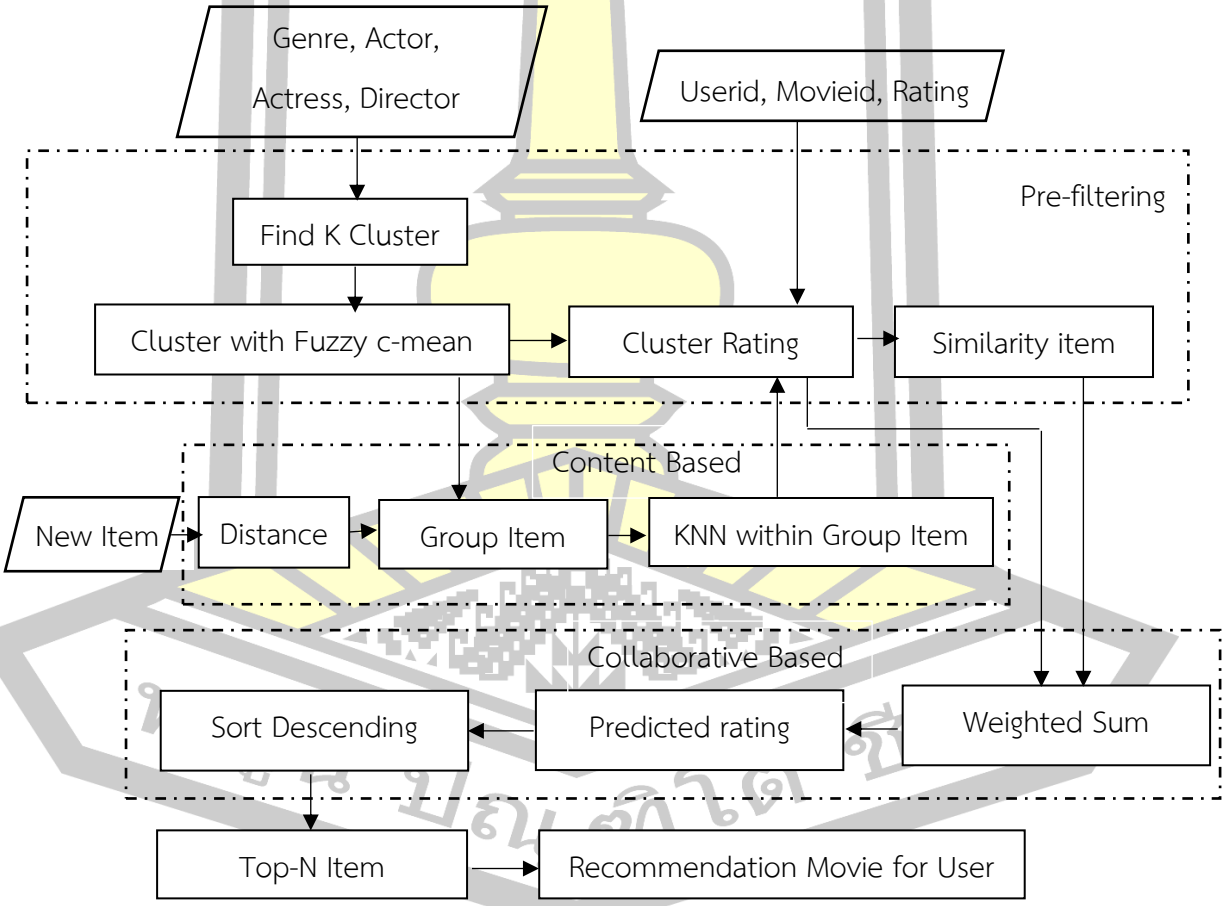
3.3.2.1 ข้อมูลนำเข้า ได้แก่ ข้อมูลประเภทของภาพยนตร์ (Genre) ข้อมูลการให้ค่าคะแนนจากผู้ใช้ในการดูภาพยนตร์ ประกอบด้วยข้อมูล รหัสผู้ใช้ (Userid) รหัสภาพยนตร์ (Movieid) และ คะแนน (Rating) จากฐานข้อมูล MovieLens ข้อมูลนักแสดงชาย (Actor) ข้อมูลนักแสดงหญิง (Actress) และข้อมูลผู้กำกับ (Director) จากฐานข้อมูล IMDB

3.3.2.2 การประมวลผลด้วย Content Based ในกรณีที่มีภาพยนตร์ใหม่ ซึ่งไม่มีข้อมูลค่าคะแนนในการดูภาพยนตร์ จะทำการนำข้อมูลของภาพยนตร์ประกอบด้วย ประเภทภาพยนตร์ นักแสดงชาย นักแสดงหญิง และผู้กำกับ วัดระยะห่างจากข้อมูลจุดศูนย์กลางของกลุ่ม (Centroid) เพื่อทำให้ทราบข้อมูลจะเข้าไปเป็นสมาชิกในกลุ่มใด เมื่อเข้าไปยังกลุ่มที่เตรียมไว้ จากนั้น

จะทำการหาค่าความใกล้เคียงของภาพยนตร์ภายในกลุ่ม ด้วยเทคนิค KNN ซึ่งจะกำหนดค่าเทรดโชว์ที่เหมาะสม และกำหนด K=12 เมื่อได้จำนวนภาพยนตร์ตามที่หาได้ แล้วไปทำตามขั้นตอนถัดไป

3.3.2.4 การประมวผลด้วย Collaborative Based เพื่อแนะนำข้อมูลภาพยนตร์ให้กับผู้ใช้ เป็นนำข้อมูลค่าคะแนน และข้อมูลความใกล้เคียงที่เตรียมไว้แล้ว ดังตาราง 10 และ ตาราง 11 นำมาหาค่าน้ำหนักรวมของข้อมูล โดยผู้วิจัยทำการหาค่าน้ำหนักรวมจำนวนตามที่ได้จากหาค่า KNN โดยทำทีละเรื่องจนครบตามจำนวนคำนวณไว้ ซึ่งจะนำข้อมูลที่มีค่าน้ำหนักมากจำนวน 10 คน ของแต่ละคนมารวมกันแล้วจัดเรียงจากมากไปน้อย แล้วนำข้อมูลภาพยนตร์ตามจำนวนที่เหมาะสมไปแนะนำ ซึ่งค่าน้ำหนักรวมของข้อมูล

3.3.2.5 การแนะนำข้อมูลภาพยนตร์ให้กับผู้ใช้ มาเลือกค่าคะแนนที่มีค่ามากที่สุด [39] [57] เพื่อแนะนำภาพยนตร์ให้กับผู้ใช้ต่อไป ทั้งนี้สามารถสรุปเป็นภาพรวมการทำงาน ดังรูปที่ 16



รูปที่ 16 ขั้นตอนการแก้ไขปัญหาภาพยนตร์ใหม่

3.4 การวัดประสิทธิภาพ

3.4.1 การวัดประสิทธิภาพด้วยค่าคลาดเคลื่อนสมบูรณ์เฉลี่ย (Mean Absolute Error : MAE) เป็นวิธีการหาค่าเฉลี่ยของความแตกต่างสมบูรณ์ระหว่างค่าพยากรณ์และค่าจริง ซึ่งหากผลการประเมินมีค่าน้อย แสดงว่าค่าที่พยากรณ์ได้มีความใกล้เคียงกับค่าจริง มีสมการ (20)

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | p_{u,i} - r_{u,i} | \quad (20)$$

เมื่อ N = จำนวนภาพยนตร์ที่แนะนำ

$p_{u,i}$ = ค่าพยากรณ์

$r_{u,i}$ = ค่าคะแนนจริง

3.4.2 ค่าความแม่นยำ (Precision) คือ เป็นการวัดความแม่นยำในการแนะนำข้อมูลภาพยนตร์ โดยค่าความแม่นยำเป็นอัตราส่วนของชิ้นข้อมูลหรือภาพยนตร์ที่ผู้ใช้ให้ความสนใจและตรงกับชิ้นข้อมูลหรือภาพยนตร์ที่แนะนำ กับจำนวนการแนะนำชิ้นข้อมูลหรือภาพยนตร์ทั้งหมด [32] [33] สมการณ์ (21)

$$\mathrm{Pr}ecision = \frac{|\{relevant\ item\} \cap \{top - N\ item\}|}{top - N\ item} \quad (21)$$

เมื่อ $relevant\ item$ คือ ชิ้นข้อมูลที่ผู้ใช้ให้ความสนใจ

$top - N\ item$ คือ ชิ้นข้อมูลที่แนะนำ

N คือ จำนวนชิ้นข้อมูลที่แนะนำ

บทที่ 4
ผลการดำเนินงาน

ในบทนี้ผู้วิจัยได้นำเสนอผลการวิเคราะห์ข้อมูลที่ได้จากการทดสอบการแนะนำข้อมูลเพื่อแก้ไขปัญหาของผู้ใช้ใหม่ ข้อมูลเบาบาง และข้อมูลภาพยนตร์ใหม่ผลการวิจัยและผลการประเมินประสิทธิภาพ ดังนี้

4.1 ผลการรวบรวมข้อมูล

การรวบรวมข้อมูลจากฐานข้อมูล MovieLens และฐานข้อมูล IMDB โดยฐานข้อมูล MovieLens เป็นชุดข้อมูลประกอบด้วยข้อมูลจำนวน 100,000 เร็คคอร์ด ผู้ใช้งาน จำนวน 943 คน และภาพยนตร์จำนวน 1,682 เรื่อง ส่วนฐานข้อมูล IMDB ข้อมูลที่นำมาใช้ประกอบด้วย ข้อมูลนักแสดงชาย (Actor) จำนวน 22,240 คน นักแสดงหญิง (Actress) จำนวน 10,942 คน และผู้กำกับ (Director) จำนวน 861 คน เมื่อทำการนำมาเป็นเมทริกซ์ดังตาราง 13

ตาราง 13 เมตริกซ์รายการข้อมูล

Matrix	Dimension
User-rating	943 x 1682
User-Occupation	943 x 21
Genre-Movie	19 x 1682
Actor-Movie	22,240 x 1682
Actress-Movie	10,942 x 1682
Director-Movie	861 x 1682

เมื่อนำข้อมูลจากฐานข้อมูล MovieLens กับฐานข้อมูล IMDB จะได้ลักษณะข้อมูลดังตาราง 14

ตาราง 14 ความสัมพันธ์ของข้อมูลจากฐาน MovieLens กับ MIDB

		Movie				
		M1	M2	M3	...	M1682
User	U1	4	0	3	0	0
	U2	0	3	2	2	0
	U3	5	3	0	0	4
	...	0	0	4	3	0
	U943	2	0	5	0	2
Genre	G1	1	0	0	0	0
	G2	1	1	0	0	1
	G3	0	1	0	0	1
	...	0	0	1	0	0
	G19	0	0	0	1	0
Actors	A1	0	1	0	1	0
	A2	0	1	0	0	0
	A3	1	0	0	0	0
	...	0	0	1	0	1
	A22240	0	0	0	1	1
Actress	B1	1	0	0	1	0
	B2	0	0	1	0	1
	B3	0	1	0	0	0
	...	0	0	1	0	1
	B10942	0	1	0	0	0
Director	D1	0	1	0	0	0
	D2	1	0	0	0	0
	D3	0	0	1	0	0
	...	0	0	0	1	0
	D861	0	0	0	0	1

จากตาราง 14 เป็นข้อมูลแสดงความสัมพันธ์หว่างผู้ใช้ (User) กับภาพยนตร์ (Movie) โดยผู้ใช้มีประสบการณ์ในการดูภาพยนตร์และให้ข้อมูลค่าคะแนน (1-5) ไว้ รวมทั้งข้อมูลที่มีนักแสดงชาย นักแสดงหญิง และผู้กำกับ ซึ่งข้อมูลทั้งหมดมีความสัมพันธ์กัน ยกตัวอย่างเช่น ผู้ใช้ U1 ผ่านการดูภาพยนตร์เรื่อง M1 และให้คะแนนค่าคะแนนเป็น 4 ซึ่งข้อมูล M1 เป็นประเภท G1 และ G2 นักแสดงชายเป็น A3 หมายถึงสมมตินักแสดงชายชื่อ A3 นักแสดงหญิงเป็น B1 หมายถึงสมมตินักแสดงหญิงชื่อ B1 และผู้กำกับเป็น D2 หมายถึงสมมติผู้กำกับชื่อ D2

4.2 ผลการจัดเตรียมข้อมูล

ผู้วิจัยรวบรวมข้อมูล โดยข้อมูลได้นำมาจาก MovieLens ชุดข้อมูลประกอบด้วยข้อมูลค่าคะแนน 100,000 เร็คคอร์ด จากการเก็บรวบรวมข้อมูลผู้ใช้งาน จำนวน 943 คนและภาพยนตร์จำนวน 1,682 เรื่อง ลักษณะข้อมูล รายละเอียดดังนี้

4.2.1 ผลการเตรียมข้อมูลของค่าคะแนนของผู้ใช้ ลักษณะข้อมูลประกอบด้วย Userid Movieid Rating ดังตาราง 15

ตาราง 15 ข้อมูลค่าคะแนน

Userid	Movieid	Rating
1	1	4
1	2	3
1	3	5
2	1	2
3	1	4

จากตาราง 15 ผู้วิจัยได้ทำการเตรียมข้อมูลโดยการแปลงตารางเป็นแบบเมทริกซ์ขนาด 943 x 1682 รายละเอียดดังตาราง 16

ตาราง 16 ผลการเตรียมข้อมูลค่าคะแนน

	Movie1	Movie2	Movie3	Movie4	Movie5
User1	5	-	5	-	4
User2	4	3	3	4	3
User3	-	-	-	4	-
User4	5	5	-	-	3
User5	5	-	5	-	2
User6	-	5	-	3	-
User7	-	-	4	3	-
User8	5	5	-	4	3
User9	-	4	4	-	-
User10	-	5	2	-	3

4.2.2 ผลการเตรียมข้อมูลประเภทภาพยนตร์ ลักษณะข้อมูลประกอบด้วย Userid Genre ดังตาราง 17

ตาราง 17 ผลการเตรียมข้อมูลประเภทภาพยนตร์

Movieid	Action	Adventure	Animation	Children	Comedy	Crime
1	1	0	1	0	0	0
2	0	1	0	0	0	0
3	0	1	1	0	0	0
4	1	0	0	0	0	0
5	0	0	0	1	0	1
6	0	0	0	0	1	0
7	0	0	0	0	0	1
8	0	1	0	1	0	1
9	0	0	0	0	0	1
10	1	0	0	1	0	1

4.2.3 ผลการเตรียมข้อมูลของผู้ใช้ ลักษณะข้อมูลประกอบด้วย Userid Age Sex Occupation ดังตาราง 18

ตาราง 18 ลักษณะข้อมูลผู้ใช้

Userid	Age	Sex	Occupation
1	24	M	technician
2	53	F	artist
3	23	M	writer
4	24	M	technician
5	33	F	doctor
6	42	M	executive
7	57	M	farmer
8	36	M	artist
9	29	M	farmer
10	53	M	doctor

จากตาราง 18 ผู้วิจัยทำการแปลงข้อมูลเพศ (Gender) จากตัวอักษรเป็นตัวเลข เช่น M=1 F=2 เปลี่ยนข้อมูลอาชีพ (Occupation) เป็นตัวเลข เช่น Technician = 1 เป็นต้น ทำให้ได้ผลการเตรียมข้อมูลผู้ใช้ดังตาราง 19

ตาราง 19 ผลการจัดเตรียมข้อมูลผู้ใช้

Userid	Age	Sex	technician	artist	Writer	doctor	executive	farmer
1	24	1	1	0	0	0	0	0
2	53	2	0	1	0	0	0	0
3	23	1	0	0	1	0	0	0
4	24	1	1	0	0	0	0	0

ตาราง 19 (ต่อ)

Userid	Age	Sex	technician	artist	Writer	doctor	executive	farmer
5	33	2	0	0	0	1	0	0
6	42	1	0	0	0	0	1	0
7	57	1	0	0	0	0	0	1
8	36	1	0	1	0	0	0	0
9	29	1	0	0	0	0	0	1
10	53	1	0	0	0	1	0	0

4.2.4 ผลการเตรียมข้อมูลนักแสดงชาย ลักษณะข้อมูลประกอบด้วย Movieid ชื่อนักแสดงชาย ดังตาราง 20

ตาราง 20 ผลการเตรียมข้อมูลนักแสดงชาย

Movieid	Alexander	Keith	Kellen	Kelly	Ken	Kenny	Kenrick
1	1	0	1	0	0	0	0
2	0	1	0	0	0	0	1
3	0	0	0	0	1	0	0

4.2.5 ผลการเตรียมข้อมูลนักแสดงหญิง ลักษณะข้อมูลประกอบด้วย Movieid ชื่อนักแสดงหญิง ดังตาราง 21

ตาราง 21 ผลการเตรียมข้อมูลนักแสดงหญิง

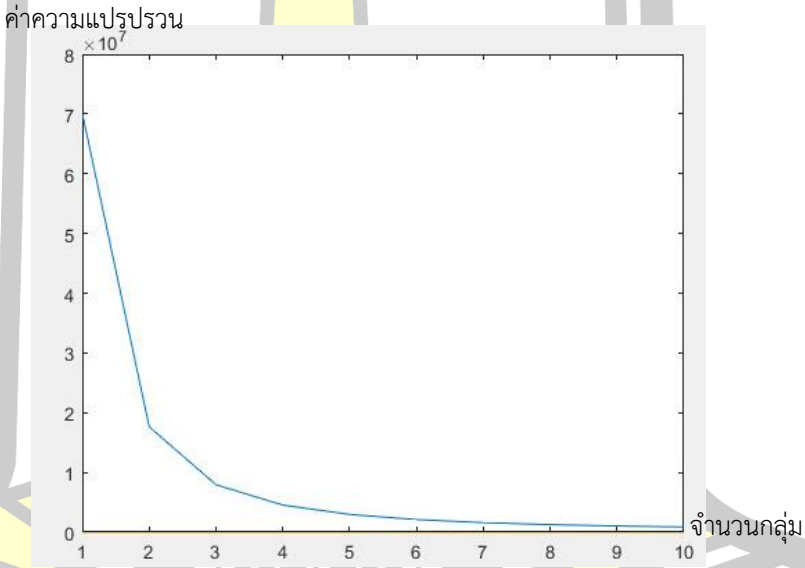
Movieid	Ancelet	Christine	Jennifer	Ancell	Luna	Ancelli	Marisa
1	0	1	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0

4.2.6 ผลการเตรียมข้อมูลผู้กำกับ ลักษณะข้อมูลประกอบด้วย Movieid ชื่อผู้กำกับ ดังตาราง 22

ตาราง 22 ผลการเตรียมข้อมูลผู้กำกับ

Movieid	Aamodt	Aamuri	Aappli	Aarden	Aarma	Aarniala	Quentin
1	0	0	1	0	0	0	0
2	1	1	0	0	0	0	0
3	0	0	1	0	0	0	0

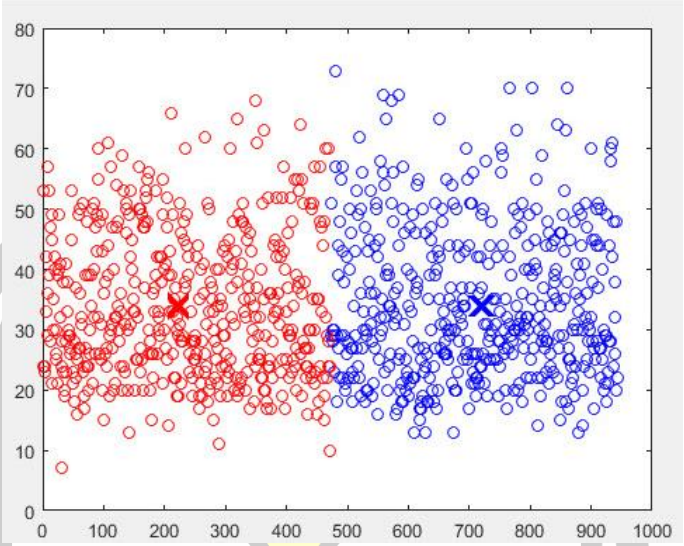
4.2.7 ผลจากการหาค่า K ก่อนการจัดกลุ่มข้อมูล เพื่อทำการหาค่า K กลุ่มที่เหมาะสมกับข้อมูลดังรูปที่ 17



รูปที่ 17 การจัดกลุ่มข้อมูล

จากรูปที่ 17 พบว่า เมื่อพิจารณาผลการจัดกลุ่มข้อมูลที่เหมาะสมที่สุด กราฟมีการหักเหที่หมายเลข 2 ฉะนั้นค่า K ในการจัดกลุ่มที่เหมาะสมคือ K=2

4.2.8 ผลการจัดกลุ่มข้อมูล เมื่อได้ค่า K ผู้วิจัยได้ทำการจัดกลุ่มข้อมูลผู้ใช้ด้วยเทคนิค Fuzzy c-mean ดังรูปที่ 18 และรูปที่ 19 เมื่อได้ข้อมูลผู้ใช้ ผู้วิจัยได้นำ Userid เป็นคีย์หลักในการค่าความสัมพันธ์กับข้อมูลตารางค่าคะแนน เพื่อทำการจัดกลุ่มข้อมูลค่าคะแนนให้เป็นไปตามข้อมูลผู้ใช้ ดังรูปที่ 18



รูปที่ 18 การจัดกลุ่มด้วย Fuzzy c-mean

จากรูปที่ 18 เป็นลักษณะการจัดกลุ่มข้อมูลผู้ใช้ด้วยเทคนิค Fuzzy c-mean โดยผู้วิจัยได้กำหนดค่า K=2 หมายถึงกำหนดกลุ่มข้อมูลเป็นจำนวน 2 กลุ่ม

24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
53	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
23	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
33	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
42	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
57	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
53	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
39	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
28	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
47	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

รูปที่ 19 การจัดกลุ่มข้อมูลผู้ใช้

1	4	0	0	0	0	0	0	0	0	2	0	0	0	4	0	0	0	0	0	0	0	0	0	4	0
2	4	0	0	0	0	0	2	4	4	0	0	4	2	0	0	0	0	0	0	0	3	3	0	0	0	0
3	0	0	0	5	0	0	0	0	5	0	3	5	0	0	0	0	0	0	0	0	0	0	3	0	3	0
4	0	0	0	0	0	0	0	0	4	0	4	5	0	0	0	0	0	0	0	0	0	0	5	0	0	0
5	0	0	0	0	0	0	0	4	0	0	2	0	0	0	5	0	0	0	0	0	0	0	0	3	3	0
6	3	0	0	5	1	0	2	4	3	0	0	5	5	0	0	0	1	0	0	0	0	4	5	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	5	0	0	0	0	0	2	0
8	0	0	0	0	0	0	0	0	0	0	0	0	1	4	4	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	5	0	0	0	0
11	5	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
12	3	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	3	0	0
13	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

รูปที่ 20 การจัดกลุ่มข้อมูลค่าคะแนน

4.2.9 ผลการจัดกลุ่มข้อมูลของประเภทภาพยนตร์ จากฐานข้อมูล MovieLens ข้อมูลนักแสดงชาย นักแสดงหญิง ผู้กำกับ จากฐานข้อมูล IMDB ผู้วิจัยการจัดกลุ่มข้อมูลผู้ใช้ด้วยเทคนิค Fuzzy c-mean เมื่อได้ข้อมูลผู้ใช้ ผู้วิจัยได้นำ Movieid เป็นคีย์หลักในการค่าความสัมพันธ์กับข้อมูลตาราง Ratting เพื่อทำการจัดกลุ่มข้อมูลค่าคะแนนให้เป็นไปตามข้อมูลของภาพยนตร์ดังรูปที่ 21

1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
4	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0
6	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
8	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
11	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
12	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
13	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0

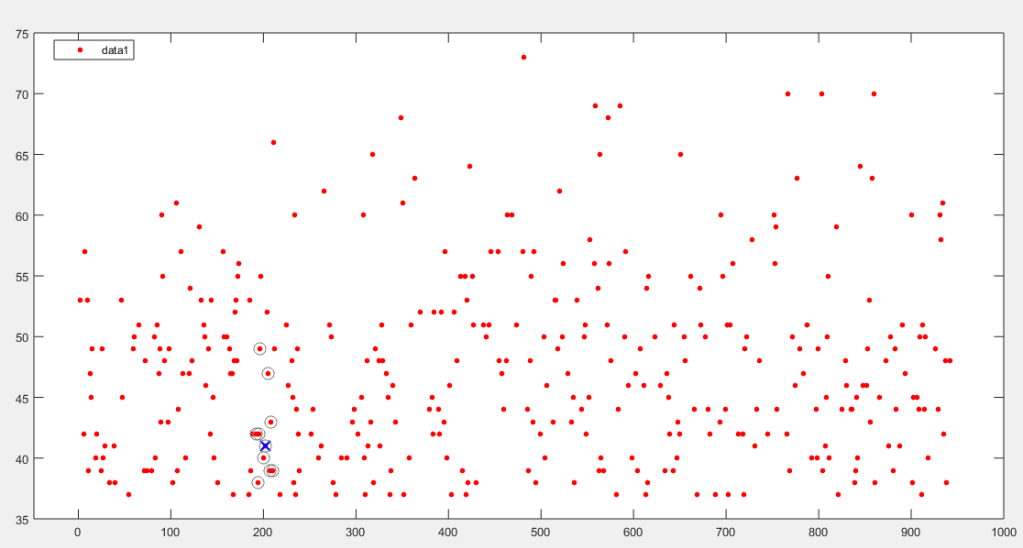
รูปที่ 21 การจัดกลุ่มข้อมูลภาพยนตร์

4.3 ผลระบบแนะนำข้อมูล

การทดลองการแนะนำข้อมูล แยกการทดลองเป็น 2 กรณี คือ กรณีผู้ใช้ใหม่ กับกรณีภาพยนตร์ใหม่ รายละเอียดดังนี้

4.3.1 ผลการทดลองการแนะนำข้อมูล กรณีผู้ใช้ใหม่ มีขั้นตอนการทำงานและผลการทดลองดังนี้

1) ผู้วิจัยนำข้อมูลผู้ใช้ใหม่เพื่อจัดเข้ากลุ่มที่เตรียมไว้ และหาค่าเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbour : KNN) ภายในกลุ่ม โดยกำหนดค่าเทรดโชว์ และค่า K=12 ได้ผลดังรูปที่ 22



รูปที่ 22 ผลการหาเพื่อนบ้านที่ใกล้ที่สุดด้วย KNN

จากรูปที่ 22 พบว่าเมื่อนำข้อมูลประกอบด้วย อายุ เพศ อาชีพ หาค่าความใกล้เคียงกับสมาชิกภายในกลุ่ม โดยกำหนดค่าเทรดโชว์ และให้ค่า K= 12 ทำให้ได้ข้อมูลที่มีความใกล้เคียงกับผู้ใช้ใหม่ที่นำมาเปรียบเทียบ

2) นำผลที่ได้จากข้อ 1 ไปทดลองหาค่าความใกล้เคียงด้วยกลุ่มค่าคะแนน ที่ละคน จำนวน 15 คน ผลการทดลองหาค่าความใกล้เคียง ในการแนะนำภาพยนตร์ให้กับผู้ใช้ดังตาราง 23

ตาราง 23 ค่าความใกล้เคียงการแนะนำภาพยนตร์

ชื่อภาพยนตร์	ค่าพยากรณ์
Contact (1997)'	4.9947
Boot, Das (1981)'	4.9913
Titanic (1997)'	4.9746
Sense and Sensibility...'	4.6775
Star Wars (1977)'	4.6447
English Patient, The ...'	4.6133
Citizen Kane (1941)'	4.5182
Full Monty, The (1997)'	4.3557
Glory (1989)'	4.1756

จากตาราง 23 ค่าความใกล้เคียงที่พยากรณ์จะมีการจัดเรียงจากค่ามากไปน้อย ซึ่งการแนะนำภาพยนตร์ให้กับผู้ใช้จำค่าพยากรณ์ที่มากที่สุด ซึ่งตามตารางจะแนะนำภาพยนตร์เรื่อง Contact

4) ผลการทดลองแนะนำภาพยนต์ กรณีผู้ใช้ใหม่

ผู้วิจัยได้ทำการทดลองกับผู้ใช้ใหม่ โดยเตรียมข้อมูลในการทดสอบจำนวน 20 เปอร์เซ็นต์ แต่ละคนไปหาค่าความใกล้เคียงกับผู้ใช้ที่มีการหาค่าโทรดโชว์กับคนที่ใกล้เคียงที่สุด ซึ่งกำหนดค่าเทรดโชว์ที่ 1.74 แต่ไม่เกินจำนวน 12 คน เพื่อวัดประสิทธิภาพด้วยค่าความแม่นยำ (Precision) กับค่าคลาดเคลื่อนสมบูรณ์เฉลี่ย (Mean Absolute Error : MAE) ได้ผลการทดลองดังตาราง 24 และตาราง 25

ตาราง 24 เปรียบเทียบผลประสิทธิภาพจากปัญหาผู้ใช้ใหม่

จำนวนข้อมูลที่ใช้ทดสอบ	Precision	
	แนะนำภาพยนต์ จำนวน 5 เรื่อง	แนะนำภาพยนต์ จำนวน 10 เรื่อง
5%	0.85	0.83
10%	0.78	0.80
15%	0.82	0.81
20%	0.85	0.84

จากตาราง 24 พบว่า เมื่อพิจารณาประสิทธิภาพจากค่าความแม่นยำ ข้อมูลที่ใช้ทดสอบ 20% การแนะนำข้อมูลภาพยนตร์ จำนวน 5 เรื่องมีความแม่นยำมากกว่า คิดเป็น 85% การแนะนำข้อมูลจำนวน 10 เรื่อง คิดเป็น 84%

ตาราง 25 ผลการประเมินประสิทธิภาพ MAE จากปัญหาผู้ใช้ใหม่

จำนวนข้อมูลที่ใช้ทดสอบ	MAE	
	แนะนำภาพยนต์ จำนวน 5 เรื่อง	แนะนำภาพยนต์ จำนวน 10 เรื่อง
5%	2.3312	2.3971
10%	2.0756	2.2649
15%	2.0778	2.2018
20%	2.1011	2.2070

จากตาราง 25 พบว่าการประเมินประสิทธิภาพ MAE ข้อมูลที่ใช้ทดสอบ 20% การแนะนำข้อมูลภาพยนตร์ จำนวน 5 เรื่องมีค่า MAE น้อยกว่าการแนะนำข้อมูลจำนวน 10 เรื่อง แสดงให้เห็นว่าการแนะนำข้อมูลภาพยนตร์ จำนวน 5 เรื่องมีประสิทธิภาพมากกว่าการแนะนำข้อมูลจำนวน 10 เรื่อง

5) ผลการทดลองการแนะนำภาพยนตร์ กรณีภาพยนตร์ใหม่

ผู้วิจัยได้ทำการทดลองกับภาพยนตร์ใหม่ โดยเตรียมข้อมูลในการทดสอบจำนวน 20 เปอร์เซ็นต์ นำข้อมูลภาพยนตร์ใหม่แต่ละเรื่องไปหาค่าความใกล้เคียงกับเรื่องที่ใกล้เคียงมีการหาค่าโทรดโชว์กับคนที่ใกล้เคียงที่สุดกำหนดค่าเทรดโชว์ที่ 12.10 แต่ไม่เกินจำนวน 12 เรื่อง เพื่อวัดประสิทธิภาพด้วยค่าความแม่นยำ (Precision) กับค่าคลาดเคลื่อนสมบูรณ์เฉลี่ย (Mean Absolute Error : MAE) ได้ผลการทดลองดังตาราง 26 และตาราง 27

ตาราง 26 เปรียบเทียบผลประสิทธิภาพจากปัญหาภาพยนตร์ใหม่

จำนวนข้อมูลที่ใช้ทดสอบ	Precision	
	แนะนำภาพยนต์กับผู้ใช้ จำนวน 5 คน	แนะนำภาพยนต์กับผู้ใช้ จำนวน 10 คน
5 %	0.87	0.85
10 %	0.82	0.81
15 %	0.84	0.84
20 %	0.87	0.86

จากตาราง 24 พบว่า เมื่อพิจารณาประสิทธิภาพจากค่าความแม่นยำ ข้อมูลที่ใช้ทดสอบ 20% การแนะนำข้อมูลภาพยนตร์ จำนวน 5 เรื่องมีความแม่นยำมากกว่า คิดเป็น 87% การแนะนำข้อมูลจำนวน 10 เรื่อง คิดเป็น 86%

ตาราง 27 ผลการประเมินประสิทธิภาพ MAE จากปัญหาภาพยนตร์ใหม่

จำนวนข้อมูลที่ใช้ทดสอบ	MAE	
	แนะนำภาพยนต์ จำนวน 5 เรื่อง	แนะนำภาพยนต์ จำนวน 10 เรื่อง
5%	2.2566	2.3721
10%	2.1763	2.2430
15%	2.0822	2.1892
20%	2.0031	2.1104

จากตาราง 27 พบว่าการประเมินประสิทธิภาพ MAE ข้อมูลที่ใช้ทดสอบ 20% การแนะนำข้อมูลภาพยนตร์ให้กับผู้ใช้ จำนวน 5 คน มีค่า MAE น้อยกว่าการแนะนำภาพยนตร์ให้กับผู้ใช้จำนวน 10 คน แสดงให้เห็นว่าการแนะนำข้อมูลภาพยนตร์กับผู้ใช้ จำนวน 5 คน มีประสิทธิภาพมากกว่าการแนะนำภาพยนตร์ให้กับผู้ใช้ จำนวน 10 คน

บทที่ 5
สรุปผล อภิปรายผล และข้อเสนอแนะ

จากการทำวิทยานิพนธ์เรื่องการเพิ่มประสิทธิภาพระบบให้คำแนะนำด้วยวิธีการผสมผสาน ซึ่งผู้วิจัยสามารถสรุปผล อภิปรายผล และข้อเสนอแนะ รายละเอียดดังนี้

5.1 สรุปผล

จากการนำข้อมูลได้นำมาจาก MovieLens ชุดข้อมูลประกอบด้วยข้อมูล ค่าคะแนน 100,000 เร็คคอร์ด จากการเก็บรวบรวมข้อมูลผู้ใช้งาน จำนวน 943 คนและภาพยนตร์จำนวน 1,682 เรื่อง และข้อมูลจาก IMDB ที่มีเรื่องสอดคล้องกับชุดข้อมูล MovieLens เพื่อมาเป็นข้อมูลในการแนะนำภาพยนตร์ให้กับผู้ใช้เพื่อใช้แก้ปัญหาผู้ใช้ใหม่ และข้อมูลภาพยนตร์ใหม่ สามารถสรุปผลการทดลอง ดังนี้

1) การแก้ไขปัญหาผู้ใช้ใหม่ ข้อมูลที่นำมาใช้ในการจัดกลุ่มคือ เพศ อายุ อาชีพ และค่าคะแนน จากฐานข้อมูล MovieLens ด้วยการเตรียมข้อมูลก่อน โดยการจัดกลุ่มด้วยวิธีการ Fuzzy c-mean ซึ่งก่อนการจัดกลุ่มเพื่อกำหนด K กลุ่มของข้อมูลนั้น ได้ทำการหาค่า K ที่เหมาะสมก่อนทำการจัดกลุ่มข้อมูลที่เหมาะสมที่สุด พบว่า ได้ค่า K=2 จากนั้นทำการจัดกลุ่ม และนำข้อมูลผู้ใช้ใหม่หาค่าเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbour) ภายในกลุ่ม โดยกำหนดค่าเทรดโชว์และกำหนดค่า K=12 จากนั้นนำข้อมูลที่ประกอบด้วยค่าคะแนน ไปหาค่าความใกล้เคียงข้อมูลของคนอื่นด้วย ผลปรากฏว่า สามารถที่จะแก้ไขปัญหาผู้ใช้ใหม่ได้อย่างมีประสิทธิภาพ โดยมีค่าความแม่นยำ ข้อมูลที่ใช้ทดสอบ 20% การแนะนำข้อมูลภาพยนตร์ จำนวน 5 เรื่องมีความแม่นยำมากกว่า คิดเป็น 85% การแนะนำข้อมูลจำนวน 10 เรื่อง คิดเป็น 84% และผลการประเมินประสิทธิภาพ MAE ข้อมูลที่ใช้ทดสอบ 20% การแนะนำข้อมูลภาพยนตร์ จำนวน 5 เรื่องมีค่า MAE น้อยกว่าการแนะนำข้อมูลจำนวน 10 เรื่อง แสดงให้เห็นว่าการแนะนำข้อมูลภาพยนตร์ จำนวน 5 เรื่องมีประสิทธิภาพมากกว่าการแนะนำข้อมูลจำนวน 10 เรื่อง

2) กรณีแก้ไขปัญหาข้อมูลใหม่ ข้อมูลที่นำมาใช้ในการจัดกลุ่มคือ ประเภทภาพยนตร์ ข้อมูลค่าคะแนน จากฐานข้อมูล MovieLens และข้อมูลนักแสดงชาย นักแสดงหญิง และผู้กำกับ จากฐานข้อมูล IMDB ด้วยการเตรียมข้อมูลก่อน โดยการจัดกลุ่มด้วยวิธีการ Fuzzy c-mean ซึ่งก่อนการจัดกลุ่มเพื่อกำหนด K กลุ่มของข้อมูลนั้น ได้ทำการหาค่า K ที่เหมาะสมก่อนทำการจัดกลุ่มข้อมูลที่

เหมาะสมที่สุด พบว่า ได้ค่า K=2 จากนั้นทำการจัดกลุ่ม และนำข้อมูลภาพยนตร์ใหม่หาค่าเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbour) ภายในกลุ่ม โดยกำหนดค่าเทรดโชว์ที่เหมาะสมและกำหนด K=12 จากนั้นนำข้อมูลที่ประกอบด้วยค่าคะแนน ไปหาค่าความใกล้เคียงข้อมูลของภาพยนตร์เรื่องอื่น ผลปรากฏว่า สามารถที่จะแก้ไขปัญหาภาพยนตร์ใหม่ได้อย่างมีประสิทธิภาพ โดยมีค่าความแม่นยำข้อมูลที่ใช้ทดสอบ 20% การแนะนำข้อมูลภาพยนตร์ จำนวน 5 เรื่องมีความแม่นยำมากกว่า คิดเป็น 87% การแนะนำข้อมูลจำนวน 10 เรื่อง คิดเป็น 86% และผลการประเมินประสิทธิภาพ MAE ข้อมูลที่ใช้ทดสอบ 20% การแนะนำข้อมูลภาพยนตร์ จำนวน 5 เรื่องมีค่า MAE น้อยกว่าการแนะนำข้อมูลจำนวน 10 เรื่อง แสดงให้เห็นว่าการแนะนำข้อมูลภาพยนตร์ จำนวน 5 เรื่องมีประสิทธิภาพมากกว่าการแนะนำข้อมูลจำนวน 10 เรื่อง

5.2 อภิปรายผล

สำหรับการแก้ไขปัญหาการแก้ไขปัญหาผู้ใช้ใหม่ และภาพยนตร์ใหม่ ข้อมูลที่นำมาใช้ในการจัดกลุ่มคือ เพศ อายุ อาชีพ ประเภทภาพยนตร์ และค่าคะแนน จากฐานข้อมูล MovieLens ข้อมูลนักแสดงชาย นักแสดงหญิง ผู้กำกับ จากฐานข้อมูล IMDB ซึ่งผลจากการทดลองด้วยการเตรียมข้อมูลก่อน โดยการจัดกลุ่มด้วยวิธีการ Fuzzy c-mean และนำข้อมูลผู้ใช้ใหม่หรือภาพยนตร์ใหม่หาค่าเพื่อนบ้านที่ใกล้ที่สุด ภายในกลุ่ม โดยกำหนดค่าเทรดโชว์ และค่า K=12 จากนั้นนำข้อมูลที่ประกอบด้วยค่าคะแนน ไปหาค่าความใกล้เคียงข้อมูลของคนอื่นด้วย ผลปรากฏว่า สามารถที่จะแก้ไขปัญหาผู้ใช้ใหม่และภาพยนตร์ใหม่ได้อย่างมีประสิทธิภาพ และเมื่อเปรียบเทียบกับการจัดเตรียมข้อมูลด้วยวิธีการ Content Based Filtering (CBF) กับ Collaborative Filtering (CF) โดยไม่มีการจัดเตรียมข้อมูลก่อนทำการประมวลผลแนะนำข้อมูล พบว่า การจัดเตรียมข้อมูลก่อนด้วยวิธีการ Fuzzy c-mean มีประสิทธิภาพมากกว่าวิธีการแบบ CBF กับ CF จะเห็นได้ว่า การจัดเตรียมข้อมูลด้วยวิธีการ Fuzzy c-mean ซึ่งเป็นการจัดกลุ่มที่สามารถจัดกลุ่มข้อมูลในการเข้าเป็นสมาชิกกลุ่มที่มีความละเอียดในการนำข้อมูลเข้าเป็นสมาชิกกลุ่ม รวมทั้งการหาค่าความใกล้เคียงภายในกลุ่ม และการหาค่าความใกล้เคียงและน้ำหนักรวมภายในกลุ่ม ส่งผลให้ประสิทธิภาพการแนะนำข้อมูลมีประสิทธิภาพมากกว่าวิธีการที่ไม่มีการจัดเตรียมข้อมูล ซึ่งสอดคล้องกับงานวิจัยของ Li และ Kim [26] ได้ศึกษาในการแก้ไขปัญหา Cold-Start ซึ่งได้นำข้อมูลคุณลักษณะของภาพยนตร์ประกอบด้วย นักแสดงชาย นักแสดงหญิง ผู้กำกับและประเภท และข้อมูลของผู้ใช้งานประกอบด้วย อายุ เพศ และอาชีพ และค่าคะแนนจากการดูภาพยนตร์ Lam และ คณะ (4) ได้นำข้อมูลประวัติของผู้ใช้งาน

ประกอบด้วย อายุ เพศ อาชีพ และค่าคะแนนคะแนนจากการดูภาพยนตร์ โดยใช้ฐานข้อมูลจาก MovieLens Li และ Kim ใช้วิธีการจัดกลุ่มจากคุณลักษณะของภาพยนตร์และผู้ใช้งาน จากนั้นทำการคำนวณค่าความใกล้เคียง ผลจากการทดลองแสดงให้เห็นว่าสามารถแก้ไขปัญหาได้ แต่สำหรับการแก้ไขปัญหาภาพยนตร์ใหม่นั้น ในการเปรียบเทียบความแม่นยำ (Precision) พบว่า วิธีการ Fuzzy c-mean มีประสิทธิภาพเท่ากับวิธีการ CBF กับ CF ทั้งนี้อาจเป็นเพราะว่า ผู้วิจัยได้นำข้อมูลภาพยนตร์ที่มีผู้ใช้เคยมีประสบการณ์ดูภาพยนตร์ในเรื่องที่เตรียมไว้เป็นจำนวนมาก จึงส่งผลให้ได้ค่าความแม่นยำไม่แตกต่างกัน

5.3 ข้อเสนอแนะ

1. ควรนำวิธีการแบบผสมผสานที่มีการเตรียมข้อมูลก่อนไปทดลองกับฐานข้อมูลอื่น เพื่อนำผลมาเปรียบเทียบประสิทธิภาพ

2. ควรขยายขีดความสามารถการประมวลผลข้อมูลไปสู่คลาวด์คอมพิวติ้งในการวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

3. ควรใช้เครื่องมือหรือภาษาอื่น ๆ ในการการประมวลผลระบบแนะนำข้อมูล ซึ่งอาจจะส่งผลให้การทดลองมีการประมวลผลที่รวดเร็วยิ่งขึ้น

บรรณานุกรม

บรรณานุกรม

[1] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, “Recommender systems survey,” Knowledge-Based Systems, vol. 46, pp. 109–132, Jul. 2013.

[2] S. K. Lee, Y. H. Cho, and S. H. Kim, “Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations,” Information Sciences, vol. 180, no. 11, pp. 2142–2155, Jun. 2010.

[3] Núñez-Valdéz ER, Cueva Lovelle JM, Sanjuán Martínez O, García-Díaz V, Ordoñez de Pablos P, and Montenegro Marín CE, “Implicit feedback techniques on recommender systems applied to electronic books,” Computers in Human Behavior, vol. 28, no. 4, pp. 1186–1193, 2012.

[4] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, “Incorporating contextual information in recommender systems using a multidimensional approach,” ACM Transactions on Information Systems, vol. 23, no. 1, pp. 103–145, Jan. 2005.

[5] M. A. Abbasi, J. Tang, H. Liu, A. Abbasi, J. Tang, and H. Liu, Trust-Aware Recommender Systems. Machine Learning book on computational trust. Chapman & Hall/CRC Press, 2014.

[6] L. Baltrunas and F. Ricci, “Experimental evaluation of context-dependent collaborative filtering using item splitting,” User Modeling and User-Adapted Interaction, vol. 24, no. 1–2, pp. 7–34, Feb. 2014.

[7] Gorgoglione M and Panniello U, “Including Context in a Transactional Recommender System Using a Pre-filtering Approach: Two Real E-commerce Applications,” presented at the 2009 WAINA ’09 International Conference, Advanced Information Networking and Applications Workshops, 2009, pp. 667–672.

[8] S. Kumar and S. Kumar, “An Approach for Recommender System by Combining Collaborative Filtering with User Demographics and Items Genres,” International Journal of Computer Applications, vol. 128, no. 13, pp. 16–24, Oct. 2015.

[9] G. Adomavicius and A. Tuzhilin, Recommendation Technologies: Survery of Current Methods and Possible Extensions. New York, USA: Stern School of Business New York University, 2004.

[10] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong, "Addressing cold-start problem in recommendation systems," presented at the Proceedings of the 2nd international conference on Ubiquitous information management and communication, ACM, 2008, pp. 208–211.

[11] Datta S, Das J, Gupta P, and Majumder S, "SCARS: A Scalable Context-Aware Recommendation System," presented at the 2015 Third International Conference, 2015, pp. 1–6.

[12] M. A. Ghazanfar and A. Prugel-Bennett, "A Scalable, Accurate Hybrid Recommender System," in 2010 Third International Conference on Knowledge Discovery and Data Mining, Phuket, 2010, pp. 94–98.

[13] H. Yildirim and M. S. Krishnamoorthy, "A random walk method for alleviating the sparsity problem in collaborative filtering," in Proceedings of the 2008 ACM conference on Recommender systems - RecSys '08, Lausanne, Switzerland, 2008, p. 131.

[14] Simon Phiip, P.B. Shola, and Abari Ovye John, "Application of content-based approach in research paper recommendation system for a digital library," International Journal of Advanced Computer Science and Applications, vol. 5, no. 10, pp. 37–40, 2014.

[15] Z. Xia, Y. Dong, and G. Xing, "Support vector machines for collaborative filtering," in Proceedings of the 44th annual southeast regional conference on - ACM-SE 44, Melbourne, Florida, 2006, p. 169.

[16] S.-H. Min and I. Han, "Recommender Systems Using Support Vector Machines," presented at the Web Engineering: 5th International Conference, ICWE 2005, Sydney, Australia, 2005, pp. 387–393.

[17] L. Baltrunas and F. Ricci, "Context-based splitting of item ratings in collaborative filtering," in Proceedings of the third ACM conference on Recommender systems - RecSys '09, New York, New York, USA, 2009, p. 245.

[18] P. G. Campos, I. Fernández-Tobías, I. Cantador, and F. Díez, “Context-Aware Movie Recommendations: An Empirical Comparison of Pre-filtering, Post-filtering and Contextual Modeling Approaches,” in E-Commerce and Web Technologies, vol. 152, C. Huemer and P. Lops, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 137–149.

[19] Y. Zheng, B. Mobasher, and R. Burke, “Context Recommendation Using Multi-label Classification,” in 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland, 2014, pp. 288–295.

[20] Tewari AS, Kumar A, and Barman AG, “Book Recommendation System Based on Combine Features of Content Based Filtering, Collaborative Filtering and Association Rule Mining,” presented at the Advance Computing Conference (IACC), 2014, pp. 500–503.

[21] L. Baltrunas, M. Kaminskas, and F. Ricci, “Best Usage Context Prediction for Music Tracks,” p. 6.

[22] M. A. Ghazanfar and A. Prugel-Bennett, “An Improved Switching Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering,” p. 10.

[23] Esfahani MH and Alhan FK, “New Hybrid Recommendation System Based On C-Means Slustering Method,” presented at the 2013 5th Conference, Information and Knowledge Technology (IKT), 2013, pp. 145–149.

[24] Qing Li and Byeong Man Kim, “Clustering approach for hybrid recommender system,” in Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003), Halifax, NS, Canada, 2003, pp. 33–38.

[25] M. Braunhofer, V. Codina, and F. Ricci, “Switching hybrid for cold-starting context-aware recommender systems,” in Proceedings of the 8th ACM Conference on Recommender systems - RecSys ’14, Foster City, Silicon Valley, California, USA, 2014, pp. 349–352.

[26] Braunhofer M, Codina V, and Ricci F, “Hybridisation Techniques for Cold-Starting Context-Aware Recommender Systems,” presented at the Proceedings of the 8th ACM Conference on Recommender systems, Foster City, Silicon Valley, California, USA: ACM, 2014, pp. 349–352.

[27] Alan Eckhardt, “Similarity of users’(content-based) preference models for Collaborative filtering in few ratings scenario,” Expert Systems with Applications, vol. 39, no. 14, pp. 11511–11516, 2012.

[28] C. Musto, “Enhanced vector space models for content-based recommender systems,” in Proceedings of the fourth ACM conference on Recommender systems - RecSys ’10, Barcelona, Spain, 2010, p. 361.

[29] Kim J, Lee D, and Chung K-Y, “Item Recommendation Based on Context-aware Model for Personalized u-healthcare Service,” Multimedia Tools and Applications, vol. 71, no. 2, pp. 855–872, 2014.

[30] H. Wu, K. Yue, X. Liu, Y. Pei, and B. Li, “Context-Aware Recommendation via Graph-Based Contextual Modeling and Postfiltering,” International Journal of Distributed Sensor Networks, vol. 11, no. 8, p. 613612, Aug. 2015.

[31] Dheeraj kumar Bokde, Sheetal Girase, and Debajyoti Mukhopadhyay, “An item-based collaborative filtering using dimensionality reduction techniques on Mahout framework,” presented at the th Post Graduate Conference for Information Technology (iPGCon-2015), Sangamner, Published in Spvryan’s International Journal of Engineering Science and Technology (SEST), 2015.

[32] Yuan-hong W and Xiao-qiu T, “Real-Time Recommender System Based on Hybrid Collaborative Filtering,” presented at the 2010 5th International Conference, 2010, pp. 1909–1912.

[33] M. Papagelis and D. Plexousakis, “Qualitative Analysis of User-based and Item-based Prediction Algorithms for Recommendation Agents,” p. 16.

[34] He L and Wu F, “A Time Context Based Collaborative Filtering Algorithm,” presented at the GRC ’09 IEEE International Conference, Granular Computing, 2009, pp. 209–213.

[35] G. Pitsilis and S. J. Knapskog, “Social Trust as a solution to address sparsity-inherent problems of Recommender systems,” p. 8.

[36] Saurabh Kumar Tiwari,Shailendra Kumar Shrivastava, “An Approach for Recommender System by Combining Collaborative Filtering with User Demographics and Items Genres,” International Journal of Computer Applications, vol. 128, no. 13, pp. 16–24, Oct. 2015.

[37] J. Gupta and J. Gadge, “Performance analysis of recommendation system based on collaborative filtering and demographics,” in 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, 2015, pp. 1–6.

[38] V. Codina, F. Ricci, and L. Ceccaroni, “Semantically-enhanced pre-filtering for context-aware recommender systems,” in Proceedings of the 3rd Workshop on Context-awareness in Retrieval and Recommendation - CaRR ’13, Rome, Italy, 2013, pp. 15–18.

[39] V. Codina, F. Ricci, and L. Ceccaroni, “Local context modeling with semantic pre-filtering,” in Proceedings of the 7th ACM conference on Recommender systems - RecSys ’13, Hong Kong, China, 2013, pp. 363–366.

[40] X. N. Lam and T. Vu, “Addressing Cold-Start Problem in Recommendation Systems,” p. 4.

[41] S. Gong, “A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering,” Journal of Software, vol. 5, no. 7, pp. 745–752, Jul. 2010.

[42] K. Choi, D. Yoo, G. Kim, and Y. Suh, “A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis,” Electronic Commerce Research and Applications, vol. 11, no. 4, pp. 309–317, Jul. 2012.

[43] Y. Chen, J. Kim, and H. S. Mahmassani, “Pattern recognition using clustering algorithm for scenario definition in traffic simulation-based decision support systems,” in 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 2014, pp. 798–803.

[44] M. Balouchestani, L. Sugavaneswaran, and S. Krishnan, “Advanced K-means clustering algorithm for large ECG data sets based on K-SVD approach,” in 2014 9th International Symposium on Communication Systems, Networks & Digital Sign (CSNDSP), Manchester, UK, 2014, pp. 177–182.

[45] Gong S, “Learning User Interest Model for Content-based Filtering in Personalized Recommendation System,” International Journal of Digital Content Technology & ITs Applications, vol. 6, no. 11, 2012.

[46] Z. Huang, D. Zeng, and H. Chen, “A Link Analysis Approach to Recommendation under Sparse Data,” New York, p. 9, 2004.

[47] Ma H, King I, and Lyu MR, “Effective Missing Data Prediction for Collaborative Filtering,” presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands: ACM, 2007, pp. 39–46.

[48] H. Luo, C. Niu, R. Shen, and C. Ullrich, “A collaborative filtering framework based on both local user similarity and global user similarity,” Machine Learning, vol. 72, no. 3, pp. 231–245, Sep. 2008.

[49] Yang X, Zhang Z, and Wang K, “Scalable Collaborative Filtering Using Incremental Update and Local Link Prediction,” presented at the Proceedings of the 21st ACM international conference on Information and knowledge management, Maui, Hawaii, USA: ACM, 2012, pp. 2371–2374.

[50] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, “Hybrid Collaborative and Content-based Music Recommendation Using Probabilistic Model with Latent User Preferences,” p. 6.

[51] J. Salter and N. Antonopoulos, “CinemaScreen Recommender Agent: Combining Collaborative and Content-Based Filtering,” IEEE Intelligent Systems, vol. 21, no. 1, pp. 35–41, Jan. 2006.

[52] Shrestha, Jenu, and Geun-Sik Jo, “Enhanced Content-Based Filtering Using Diverse Collaborative Prediction for Movie Recommendation,” presented at the 2009 ACIIDS 2009 First Asian Conference, Intelligent Information and Database Systems, 2009, pp. 132–137.

[53] Wang Q, Yuan X, and Sun M, “Collaborative Filtering Recommendation Algorithm Based on Hybrid User Model,” presented at the 2010 Seventh International Conference, Fuzzy Systems and Knowledge Discovery (FSKD), 2010, pp. 1985–1990.

[54] S. K. Shinde and U. V. Kulkarni, “Hybrid Personalized Recommender System Using Fast K-medoids Clustering Algorithm,” Journal of Advances in Information Technology, vol. 2, no. 3, pp. 152–158, Aug. 2011.

[55] W. Yuan-hong and T. Xiao-qiu, “A real-time recommender system based on hybrid collaborative filtering,” in 2010 5th International Conference on Computer Science & Education, Hefei, China, 2010, pp. 1909–1912.

[56] Gupta A, Shivhare H, and Sharma S, “Recommender system using fuzzy c-means clustering and genetic algorithm based weighted similarity measure,” presented at the 2015 International Conference, Computer, Communication and Control (IC4), 2015, pp. 1–8.

[57] A. Spaeth and M. C. Desmarais, “Combining Collaborative Filtering and Text Similarity for Expert Profile Recommendations in Social Websites,” in User Modeling, Adaptation, and Personalization, vol. 7899, S. Carberry, S. Weibelzahl, A. Micarelli, and G. Semeraro, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 178–189.

ประวัติผู้เขียน

ชื่อ	นายธรรมนูญ ปัญญาทิพย์
วันเกิด	วันที่ 13 ตุลาคม พ.ศ. 2519
สถานที่เกิด	อำเภอแกดำ จังหวัดมหาสารคาม
สถานที่อยู่ปัจจุบัน	340 หมู่ 12 ตำบลเกิ้ง อำเภอเมือง จังหวัดมหาสารคาม 44000
ตำแหน่งหน้าที่การงาน	อาจารย์
สถานที่ทำงานปัจจุบัน	คณะวิทยาศาสตร์และเทคโนโลยีสุขภาพ มหาวิทยาลัยกาฬสินธุ์ ตำบลสงเปลือย อำเภอนามน จังหวัดกาฬสินธุ์ 46230
ประวัติการศึกษา	พ.ศ. 2538 มัธยมศึกษาตอนปลาย โรงเรียนเหล่าล้อวิทยาคม อำเภอศรีสมเด็จ จังหวัดร้อยเอ็ด พ.ศ. 2542 ปริญญาบริหารธุรกิจบัณฑิต (บธ.บ.) สาขาคอมพิวเตอร์ธุรกิจ มหาวิทยาลัยมหาสารคาม พ.ศ. 2552 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สื่อนฤมิต มหาวิทยาลัยมหาสารคาม พ.ศ. 2562 ปริญญาปรัชญาดุษฎีบัณฑิต (ปร.ด.) วิทยาการคอมพิวเตอร์ มหาวิทยาลัยมหาสารคาม